AD-A052 427 TEXAS TECH UNIV LUBBOCK INST FOR ELECTRONICS SCIENCE F/G 9/3
SEMIANNUAL REVIEW OF RESEARCH UNDER THE ASSOCIATE JOINT SERVICE--ETC(U)
OCT 77 R SAEKS, K S CHAO, S R LIBERTY N00014-76-C-1136

UNCLASSIFIED NL

1 OF 2
ADA
052427

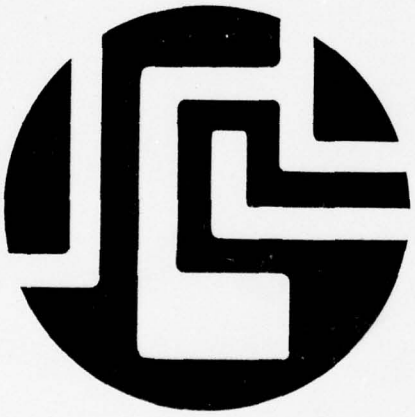
(12)

SEMIANNUAL REVIEW OF RESEARCH

under the

ASSOCIATE JOINT SERVICES ELECTRONICS PROGRAM

October 1977



Institute for Electronics Science

TEXAS TECH UNIVERSITY
Lubbock, Texas 79409

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) (6) SEMIANNUAL REVIEW OF RESEARCH UNDER THE ASSOCIATE JOINT SERVICES ELECTRONICS PROGRAM.	5. TYPE OF REPORT & PERIOD COVERED (9) INTERIM rept.
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) (10) R. SAEKS, K.S. CHAO, S.R. LIBERTY, D. GUSTAFSON, J. WALKUP, AND T. NEWMAN	8. CONTRACT OR GRANT NUMBER(s) (15) N0014-76-C-1136
9. PERFORMING ORGANIZATION NAME AND ADDRESS TEXAS TECH UNIVERSITY INSTITUTE FOR ELECTRONIC SCIENCE LUBBOCK, TEXAS 79409	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 122105
11. CONTROLLING OFFICE NAME AND ADDRESS (11) 30 Oct 77	12. REPORT DATE 10/30/77
	13. NUMBER OF PAGES 184 + iv
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) OFFICE OF NAVAL RESEARCH 800 N. QUINCY AVE. ARLINGTON, VA. (12) 173p.	15. SECURITY CLASS. (of this report) UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE- DISTRIBUTION UNLIMITED

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

FAULT ANALYSIS
COMPUTER-AIDED DESIGN
STOCHASTIC CONTROL AND ESTIMATION
DECENTRALIZED CONTROL
MATHEMATICAL SYSTEM THEORY
OPTICAL NOISE
PATTERN RECOGNITION

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

THE FOLLOWING REPORT REPRESENTS THE SECOND SEMIANNUAL REVIEW OF RESEARCH CONDUCTED UNDER THE AUSPICES OF THE ASSOCIATE JOINT SERVICES ELECTRONICS PROGRAM AT THE INSTITUTE FOR ELECTRONIC SCIENCE AT TEXAS TECH UNIVERSITY. SPECIFIC TOPICS COVERED INCLUDE FAULT ANALYSIS, COMPUTER-AIDED DESIGN, STOCHASTIC CONTROL AND ESTIMATION, DECENTRALIZED CONTROL, MATHEMATICAL SYSTEM THEORY, OPTICAL NOISE, AND PATTERN RECOGNITION.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED LB

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

410 354

SEMIANNUAL REVIEW OF RESEARCH

under the

ASSOCIATE JOINT SERVICES ELECTRONICS PROGRAM

at the

INSTITUTE FOR ELECTRONIC SCIENCE

TEXAS TECH UNIVERSITY

October 1977

Lubbock, Texas 79409

PREFACE

The following report represents the second semiannual review of research conducted under the auspices of the Associate Joint Services Electronics Program at the Institute for Electronics Science at Texas Tech University. Specific topics covered include, fault analysis, computer-aided design, stochastic control and estimation, decentralized control, mathematical system theory, optical noise, and pattern recognition.

Table of Contents

RESEARCH

on

FAULT ANALYSIS

H.S.M. Chen and R. Saeks

DEPARTMENT OF ELECTRICAL ENGINEERING

TEXAS TECH UNIVERSITY

Abstract

A search algorithm for the solution of the fault diagnosis equations arising in linear time invariant analog circuits and systems is presented. By exploitation of Householder's formula an efficient algorithm whose computational complexity is a function of the number of system failures rather than the number of system components is obtained.

Introduction

Conceptually, the fault analysis problem for an analog circuit or system amounts to the measurement of a set of externally accessible parameters of the system from which one desires to determine the internal system parameters or equivalently[1] locate the failed components as illustrated in Figure 1. Here, the

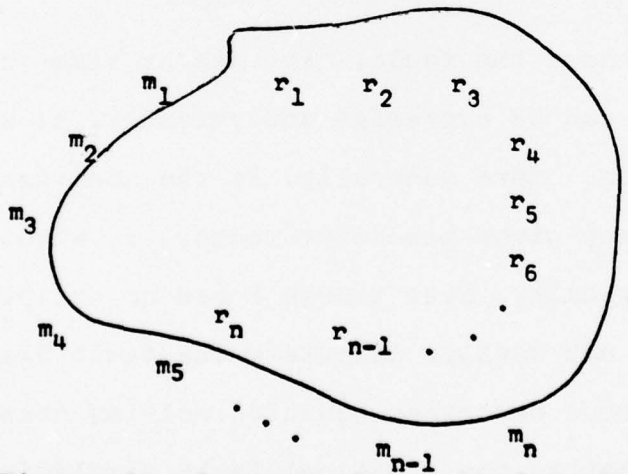


Figure 1. Conceptual Model of Fault Diagnois Problem.

measurements, m_i, may represent data taken at distinct test points or alternatively, data taken at a fixed test point under different stimuli.

[1]Since the problem of determining the values of the failed components is usually straightforward, once the failures have been located, the exact determination of all internal component parameters is essentially equivalent to the problem of "simply" locating the failed components.

Similarly, the r_i represent parameters characterizing the various internal system components. Here, a single parameter may characterize an entire component, say a resistance, capacitance or inductance. Alternatively, a component may be represented by several parameters: the h-parameters of a transistor, the poles and gain of an op-amp, etc. In general, one models a system component by the minimum number of parameters which will allow the failure to be isolated up to a "shop replaceable assembly" with all "allowed" system failures manifesting themselves in the form of some parameter change.

To solve the fault diagnosis problem, one then measures $m = \text{col}(m_i)$ and solves a nonlinear algebraic equation

$$1. \qquad m = F(r)$$

for $r = \text{col}(r_i)$ to diagnose the fault. For linear time-invariant systems the function F can be expressed analytically, as we shall see in the following section. More generally, in the nonlinear case, one can evaluate $F(r)$ for any given parameter vector, r, with a simulator, and thus solve 1. numerically, even though F has no analytic expression.

Although one does not usually formulate the fault diagnosis problem in terms of the above described equation solving notation, this formulation is equivalent to the classical fault simulation concept.[9] Indeed, fault simulation is simply a search algorithm for solving 1. Here, one precomputes $\hat{m} = f(\hat{r})$ for each allowable[1] faulty parameter vector $\hat{r}$ and then compares the measured m with the simulated $\hat{m}$'s, stored in a fault dictionary, to solve equation 1.

[1] By allowable faults we mean all possible parameter vectors, $\hat{r}$, which satisfy a specified set of fault hypotheses. These typically restrict the maximum number of component parameters which are simultaneously out of tolerence and the type of failure (open circuit, short circuit, small change, etc.)

Although the above described approach to fault simulation has been successful[1] when applied to digital system, there is considerable question surrounding its applicability to analog circuits and systems[1]. The problem here is two-fold. First, rather than simply failing as a one or zero, an analog parameter has a continuum of possible failures. Secondly, unlike a digital system wherein a component is either good or bad, in an analog system, a component parameter is either in tolerence or out of tolerence. As such, for each hypothesized failure, it may prove necessary to do an entire family of Monte Carlo simulations in which the values of the good components are randomly chosen within their tolerence limits. Although, at the present time we have insufficient practical experience to determine the precise number of fault simulations required for analog fault diagnosis, it is estimated that the number of simulations required for an analog system will exceed the number of simulations required for a digital system of similar complexity by a factor ranging between two and six order of magnitude.[1] As such, the fault simulation concept which has proven to be so successful for a digital system may not be applicable in the analog case.

As an alternative to fault simulation, one may adopt one of the more classical equation solving algorthms for the solution of 1.[2,3] Here, one first measures m and on the basis of this measurement, makes an initial guess r^o (usually taken to be nominal parameter vector) at the solution of the equations. One then evaluates $m^o = F(r^o)$ and compares it with m. If $m^o = m$, r^o is the solution to the fault diag-

[1] Most industrial users of ATE obtain satisfactory fault detection in digital circuits via fault simulation techniques but require guided probe techniques in addition to the fault dictionary data for fault diagnosis (isolation).

nosis equation. If not, one makes a new "educated" guess at the solution, r^1, (usually based on the deviation between m and m^o) and repeats the process by evaluating $m^1 = F(r^1)$ and comparing it with m. Hopefully, sequence of component parameter vectors, r^i, and simulated data vectors, $m^i = F(r^i)$, is obtained which "quickly" converges to r and m, respectively. Since the evaluation of $F(r^i)$ is essentially equivalent to the simulation of the system with the faulty parameter values, r^i, this technique is really another form of fault simulation. In this case, however, one simulates the system after the data vector has been measured and uses this data to make an educated guess at a (hopefully) small number of parameter vectors at which the system should be simulated. As such, the approach has been termed <u>simulation after test</u>[1] to distinguish it from the classical approach wherein all <u>simulation</u> is done <u>before test</u>.[1]

At the time of this writing, both approaches are under study[1], neither of which have been shown to be superior. Fault "simulation after test" requires that one include an efficient simulator in the ATE itself, which can be used for on-line computation of $m^i = F(r_i)$ after the UUT has been measured. On the other hand, simulation after test eliminates the requirement of searching a large fault dictionary for the (approximate) data matches required by "simulation before test." In addition, the complex ATPG requirement for "simulation before test" is eliminated.

To make "simulation after test" feasible, however, an efficient equation solving algorithm is required to obtain convergence of the r^i sequence in a reasonable amount of time. Moreover, since "real world" failures in analog circuits and systems often take the form of open and

short circuited components or large parameter deviations from nominal the classical perturbational algorithms a-la Newton-Raphson are inapplicable. Fortunately, in the context of the fault diagnosis problem, one can reasonable assume that relatively few component parameters have failed. As such, even though it is not valid to assume that $r-r^o$ (the deviation of r from nominal) is small in norm, it is reasonable to assume that it is small in "rank." The purpose of the present paper is to formulate a search algorithm for the solution of the fault diagnosis equations which exploits such an assumption.

In the following section, the explicit form for the fault diagnosis equations arising in linear time-invariant circuits and systems is derived.[3] Householder's formula[4] is then used to exploit the special form of these equations in combination with an assumption that r differs from r^o in relatively few coordinants to formulate a search algorithm for the solution of the fault diagnosis equations in which the computational complexity of the simulation process is a function of the number of the failures rather than the number of components. This algorithm is based on a similar algorithm suggested by Temes[5] for "simulation before test" and a large-change sensitivity algorithm first given by Leung and Spence.[6] Finally, examples of the application of the algorithm to active and passive circuits are presented and a study of the robustness of the algorithm to deviations of the "good" components from their nominal values is presented.[7]

Explicit Form of the Fault Diagnosis Equations

In the case of a linear time-invariant circuit or system, the fault diagnosis equations discussed abstractly in the previous section, may be expressed explicitally in analytical form. Indeed, it is the ex-

plicit nature of this form which makes our simplified solution algorithm possible. Since the fault diagnosis equations deal with the relationship between the externally measureable system parameters, m, and the internal component parameters, r, we adopt a "component connection model" as the starting point for the derivation of the fault diagnosis equations.[8] This is one of several commonly employed large scale system models in which the components and connections in a circuit or system are modeled by distinct equations, thereby permitting one to explicitally deal with the relationship between the individual component parameters and the composite system parameters.

Since the present study is resticted to linear time-invariant systems, we assume that each component is characterized by a transfer function matrix which is dependent on the potentially variable component parameters, $Z_i(s,r)$. For the classical RLC components $Z_i(s,r)$ may take the form R, Ls, or 1/sC for the case of a resistor, inductor, or capacitor, respectively. More generally, one may model an op-amp by the transfer function $k/(s-p_1)(s-p_2)$ where the parameter vector, r, now represents the three potentially variable component parameters; k, p_1, and p_2; or a delay by ke^{sT}, etc. Although the symbol Z is used (for historical reasons), the components are not assumed to be represented by impedance matrices. Indeed, hybrid models are used in most of our examples. For the purpose of analysis, it is assumed that all faults manifest themselves in the form of changes, possibly catastophic, in the parameter vector, r, with the frequency characteristics of the components unchanged. Although not universal, this fault hypothesis covers the most commonly encountered situations and subsumes the common industrial practice of assumming that all failures in analog circuits

and systems take the form of open and short circuited components.[1]

Our system components are thus characterized by a set of simultaneous equations

2. $$b_i = Z_i(s,r)a_i \qquad i = 1,2, \ldots, n$$

where a_i and b_i denote the component input and output vectors, respectively. For notational brevity, these component equations may be combined into a single block diagonal matrix equation

3. $$b = Z(s,r)a$$

where $b = \text{col}(b_i)$, $a = \text{col}(a_i)$ and $Z(s,r) = \text{diag}(Z_i(s,r))$. Although 3. is written as a single equation, it is important to remember that it represents a set of decoupled, simultaneous equations, in which $Z(s,r)$ is block diagonal. Indeed, we will exploit this fact in the application of Householder's formula.

Although there are many ways to represent the connection in a circuit or system; say a block diagram, linear graph or signal flow graph, any such representation is simply a graphical means for displaying a set of connection equations: Kirchoff laws, adder equations, etc. As such, for our component connection model we adopt[8] a purely algebraic connection model in which the connection equations are displayed explicitally without the intermediary of some kind of graphical connection diagram. This takes the form

4. $$a = L_{11}b + L_{12}u$$

$$y = L_{21}b + L_{22}u$$

where u and y represent the vectors of accessible inputs and outputs

which are available to the test system. In simple systems, the connection matrices, L_{ij}, are usually obtainable by inspection, whereas, in more complex systems, computer codes have been developed for their derivation.[13] Moreover, they are assured to exist in all but the most pathalogical systems.[8]

It is the pair of simultaneous matrix equations 3 and 4 which are termed the component connection model. By combining equations 3 and 4 to eliminate the component input and output variables, a and b, one may derive[3,8] an expression for the transfer function matrix observable by the test system between the test input and output vectors, u and y, obtaining

$$5. \qquad S(s,r) = L_{22} + L_{21}(1 - Z(s,r)L_{11})^{-1}Z(s,r)L_{12}$$

where

$$6. \qquad y = S(s,r)u$$

For a linear time-invariant system the transfer function S(s,r) is a complete description of the measurable data about the UUT available to the test system. Moreover, being rational it is completely determined by its value at a finite number of frequencies. As such, without loss of generality we may take our vector of measured data to be of the form

$$7. \qquad m = \text{col}[S(s_1,r), S(s_2,r), \ldots, S(s_k,r)]$$

The fault diagnosis equations then take the form

$$8 \qquad m \triangleq \begin{bmatrix} S(s_1,r) \\ S(s_2,r) \\ \cdot \\ \cdot \\ \cdot \\ S(s_k,r) \end{bmatrix} = \begin{bmatrix} L_{22}+L_{21}(1-Z(s_1,r)L_{11})^{-1}Z(s_1,r)L_{12} \\ L_{22}+L_{21}(1-Z(s_2,r)L_{11})^{-1}Z(s_2,r)L_{12} \\ \cdot \\ \cdot \\ \cdot \\ L_{22}+L_{21}(1-Z(s_k,r)L_{11})^{-1}Z(s_k,r)L_{12} \end{bmatrix} \triangleq F(r)$$

In the present context we will assume that s_1, s_2, ..., s_k represent sufficiently many frequencies to permit the fault diagnosis equations to be solved. Indeed, algorithms for determining such a set of frequencies when they exist are given in references 10,11, and 12. The problem at hand is the development of an efficient algorithm for the solution of these fault diagnosis equations.

Householder's Formula and the Search Algorithm

Given the explicit form of fault diagnosis equations of 8, it is apparent that the vast majority of the computation required for the simulation of F(r), either before or after test, is the inversion of the family of matrices; $(1 - Z(s_i,r)L_{11})$, $i = 1,2, \ldots, k$. Fortunately, given the assumption that relatively few components have failed, i.e. that r differs from its nominal value, r^o, in only a small number of coordinents, Householder's formula[4] may be invoked to compute $(1-Z(s_i,r)L_{11})^{-1}$ in terms of $(1 - Z(s_i,r^o)L_{11})^{-1}$ together with the inversion of a small dimensional matrix. More precisely, if A, B, C, and D are given matrices of dimension nxn, nxn, nxp, and pxn, respectively, where

9. $$A = B + CD$$

then

10. $$A^{-1} = [1 - B^{-1}C(1 + DB^{-1}C)^{-1}D]B^{-1}$$

As such, once B^{-1} is known, one may compute the inverse of the nxn matrix, A, in terms of B^{-1} and the inverse of the pxp matrix $(1 + DB^{-1}C)$. This technique has been used effectively for large change sensitivity analysis[6] and has recently been suggested by Temes for application to fault simulation.[5] This is achieved by exploiting the block diagonal character of Z(s,r). Thus if r differs from r^o in q coordinents Z(s,r) will differ from $Z(s,r^o)$ only in the pxp block composed of components which are effected by the faulty parameters[1]. If the rows and columns of Z(s,r) are re-ordered so that this block appears in the upper left corner of Z(s,r) then,

11. $$Z(s_i r) = Z(s_i, r^o) + \begin{bmatrix} \Delta & 0 \\ \hline 0 & 0 \end{bmatrix}$$

where Δ is pxp and Z(s,r) is nxn. We then have

12. $$(1 - Z(s_i,r)L_{11}) = (L - Z(s_i,r^o)L_{11}) + \begin{bmatrix} -\Delta(s_i,r) \\ \hline 0 \end{bmatrix} L_{11}^p$$

where L_{11}^p denotes the upper (after reordering) p rows of L_{11}. Finally, an application of Householder's formula yields

[1] Here, p is the sum of the dimensions of all the blocks of Z(s,r) which are dependent on the q coordinents in which r differs from r^o. Typically, $q \leq p$ with the exact relationship depending the block sizes.

13. $$(1-Z(s_i,r)L_{11})^{-1}$$

$$= \left[1-(1-Z(s_i,r^o)L_{11})^{-1}\begin{bmatrix}-\Delta(s_i,r)\\ \hline 0\end{bmatrix}\right.$$

$$\left.\left(1+L_{11}^p(1-Z(s_i,r^o)L_{11})^{-1}\begin{bmatrix}-\Delta(s_i,r)\\ \hline 0\end{bmatrix}\right)^{-1}L_{11}^p\right](1-Z(s_i,r^o)L_{11})^{-1}$$

Although quite complex, the only matrix major computation required for the inversion of $(1 - Z(s_i,r)L_{11})$ via 13 is the inversion of the pxp matrix in parentheses. As such, as long as the number of faulty parameter values remains small, equation 13 represents an extremely efficient means of carrying out a large number of fault simulations with relatively little computational capacity. Although Temes originally suggested the technique in the context of a "simulation before test" algorithm, the above application of Householder's formula is ideally suited for "simulation after test", wherein, it reduces the computational requirements for the simulation process to well within the capabilities of the minicomputers usually found in modern ATE.

Although Householder's formula yields an efficeint means for solving the fault diagnosis equations once the faulty parameters have been determined, it remains to locate the set of fault parameters. Fortunately, the efficiency of the solution algorithm based on Householder's formula is such that one can justify a search through "all" allowable sets of faulty parameters to locate the actual failures. Indeed, if we denote the "reduced fault diagnosis equations" in which all component values are assumed to be nominal except for q specified parameters; $r_{i(1)}$, $r_{i(2)}$, ..., $r_{i(q)}$; by

$F_{i(1)', i(2),..., i(q)}$ then the equation

14. $$m = F_{i(1), i(2), \ldots i(q)}(r_{i(1)}, r_{i(2)}, \cdots r_{i(q)})$$

will have a solution if and only if the faulty parameter values are among the $r_{i(1)}, r_{i(2)}, \cdots r_{i(q)}$. As such, if one attempts to solve 14 for each allowable family of faulty parameters, the actual fault will be indicated by the existence of a solution to the equation.

Although such a search algorithm might at first seem to be highly inefficient, when one observes that with the aide of Householder's formula, the evaluation of $F_{i(1), i(2), \ldots, i(q)}$ requires only the inversion of pxp ($p \simeq q$) matrix it is seen that this is not the case. Moreover, if one searches for the most likely failures first, relatively few equations need be solved in practice. In actual implementation in a "simulation after test" algorithm, one can readily search through all possible combinations of one, two, or three simultaneous failures, and commonly encountered combinations of larger numbers of failures, thus locating the far majority of failures in a reasonable amount of ATE time.

An alternative formulation of the search algorithm which alleviates the numerical difficulties associated with the attempt to solve a set of equations which may not have a solution (as is the case whenever one attempts to solve 14 with the wrong choice of faulty parameters) is to employ an optimization algorithm, rather than an equations solver, to minimize

15. $J_{i(1),\, i(2),\, \ldots\, i(q)}(r_{i(1)}, r_{i(2)}, \cdots r_{i(q)})$

$$= \| m - F_{i(1),\, i(2),\, \ldots\, i(q)}(r_{i(1)}, r_{i(2)}, \cdots r_{i(q)}) \|^2$$

Since 15. has a zero minimum if and only if 14. has a solution a search through the minimization of 15. for all allowable sets of faulty parameters will also locate the faulty parameters (indicated by a zero minimum).

Examples

As a first example, consider the LC filter shown in Figure 2. for which

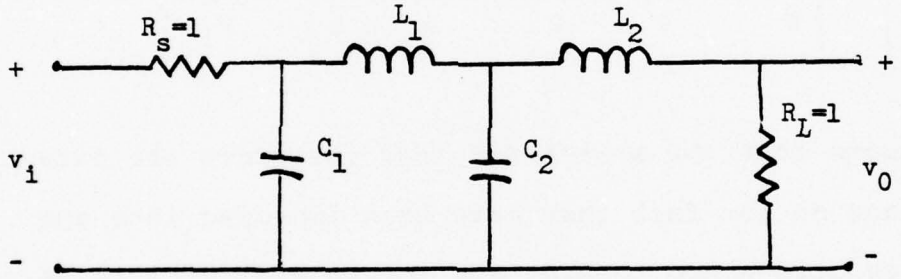


Figure 2. LC Filter.

the component connection model takes the form

$$\begin{bmatrix} v_{c1} \\ i_{L1} \\ v_{c2} \\ i_{L2} \end{bmatrix} = \begin{bmatrix} 1/S_{c1} & 0 & 0 & 0 \\ 0 & 1/S_{L1} & 0 & 0 \\ 0 & 0 & 1/S_{c2} & 0 \\ 0 & 0 & 0 & 1/S_{L2} \end{bmatrix} \begin{bmatrix} i_{c1} \\ v_{L1} \\ i_{c2} \\ v_{L2} \end{bmatrix} \qquad 16$$

16.

and

$$\left[\begin{array}{c} i_{c1} \\ v_{L1} \\ i_{c2} \\ v_{L2} \\ \hline v_o \end{array}\right] = \left[\begin{array}{cccc:c} -1 & -1 & 0 & 0 & 1 \\ 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ \hdashline 0 & 0 & 0 & 1 & 0 \end{array}\right] \left[\begin{array}{c} v_{c1} \\ i_{L1} \\ v_{c2} \\ i_{L2} \\ \hline v_i \end{array}\right] \qquad 17.$$

Since we assume that the source and load resistors are external to the filter and do not fail they have been imbedded into the connection equations and thus do not appear explicitally as components. The filter components are assumed to have the nominal values

$$C_1 = 10,\ L_1 = 20,\ C_2 = 30, \text{ and } L_2 = 40 \qquad 18.$$

and it is assumed that no more than one component fails at a time (though the failure may be catastrophic). Our "simulation after test" fault diagnosis algorithm then requires that we minimize $J_1(C_1)$, $J_2(L_1)$, $J_3(C_2)$, and $J_4(L_2)$. The performance measure with zero minimum then represents the failed component with the minimizing value for that performance measure representing the value of the failed component. All other component values must then be nominal (since it is assumed that only one component

fails). Note: the minimizing value for the non-zero J_i's does not correspond with the correct component values for those components.

This filter was simulated with each of its four components out of tolerence (by as much as 100 percent) with the search algorithm being applied to the simulated data. Since only one parameter is assumed to fail at a time and Z(s,r) is diagonal each of the four required minimizations was carried out by purely scalar operations using a Golden Section search. In all four cases the fault was correctly located with the faulty parameter value being determined "exactly." The resultant data is summarized in Table 1. Note: in each case the minimum value for J_i for the faulty component is at least three orders of magnitude lower than the minimum value J_i for any non-faulty component. As such, the failure is easily located and one can expect the algorithm to remain viable in the face of numerical and or approximation error.

As a more sophisticated example, consider the one stage transistor amplifier of Figure 3 and its wide band equivalent circuit shown in Figure 4. Note that the parallel resistors, R_a and R_b, appearing in this model have been lumped together into a single resistance, R_s, since it is clearly impossible to distinquish between failures in these two components from external measurements.

18.

$$\begin{bmatrix} V_{c1} \\ V_{rx} \\ V_{r\pi} \\ V_{cu} \\ V_{c2} \\ I_{RS} \\ I_{Re} \\ I_{c\pi} \\ I_{Ce} \\ I_{gm} \\ I_{Rc} \\ I_{RL} \end{bmatrix} = \begin{bmatrix} 1/c_{1S} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & r_x & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & r_\pi & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/c_{uS} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/C_{2S} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/R_S & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/R_e & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & C_{\pi S} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & C_{eS} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & g_m & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/R_c & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/R_L \end{bmatrix} \begin{bmatrix} I_{c1} \\ I_{rx} \\ I_{r\pi} \\ I_{cu} \\ I_{C2} \\ V_{RS} \\ V_{Re} \\ V_{C\pi} \\ V_{Ce} \\ V_{gm} \\ V_{Rc} \\ V_{RL} \end{bmatrix}$$

and

19.

$$\begin{bmatrix} I_{c1} \\ I_{rx} \\ I_{r\pi} \\ I_{cu} \\ I_{c2} \\ V_{RS} \\ V_{Rc} \\ V_{c\pi} \\ V_{ce} \\ V_{gm} \\ V_{Rc} \\ V_{RL} \\ \hline V_o \end{bmatrix} = \left[\begin{array}{cccccccccccc|c} 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ -1 & -1 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline -1 & -1 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right] \begin{bmatrix} V_{c1} \\ V_{rx} \\ V_{r\pi} \\ V_{cu} \\ V_{c2} \\ I_{RS} \\ I_{R_c} \\ I_{c\pi} \\ I_{c_e} \\ I_{gm} \\ I_{Rc} \\ I_{RL} \\ \hline V_i \end{bmatrix}$$

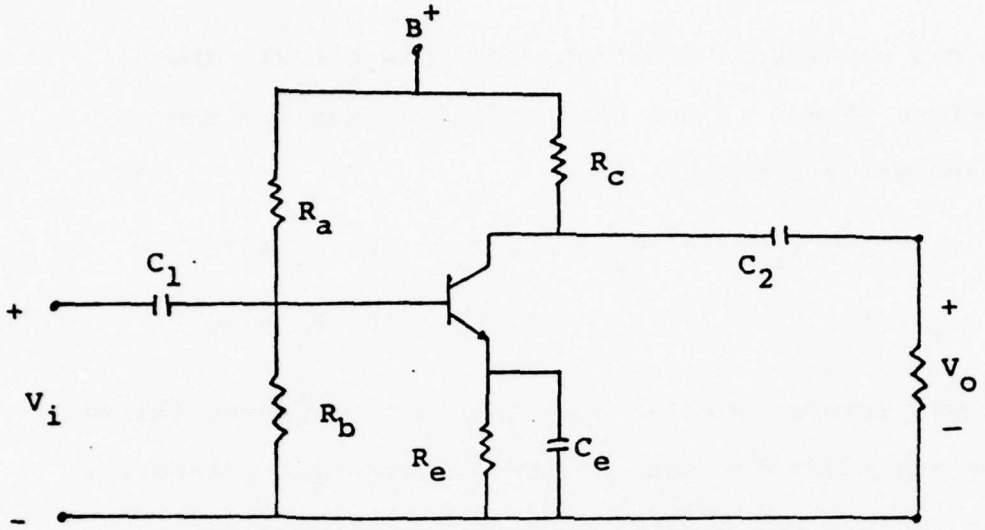


Figure 3. One stage transistor amplifier.

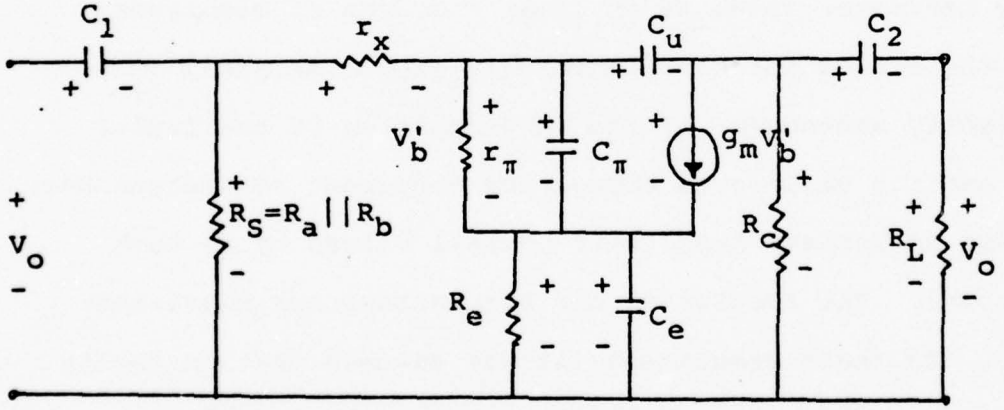


Figure 4. Amplifier equivalent circuit

The component and connection equations for this circuit are given by equations 18 and 19 and the nominal values for the component parameters are taken to be

$$C_1 = 20,\ r_x = 10,\ r_\pi = 40,\ C_u = 25,\ C_2 = 20,\ R_s = 75 \quad (20)$$
$$R_e = 30,\ C_\pi = 15,\ C_e = 10,\ g_m = 10,\ R_c = 10,\ R_l = 20$$

As before, it was assumed that no more than one component failed and $J_i(r_i)$ was minimized for each of the 12 component parameters. Once again the failure was clearly located by the smallest minima with accurate determination of the faulty parameter value. Indeed, in each case, the minimum value of $J_i(r_i)$ for the faulty parameter value is at least 5 orders of magnitude less than the minima for the remaining $J_i(r_i)$. As such, there is no ambiguity whatsoever in the determination of the faulty component and its value even though the component parameters have been allowed to deviate from their nominal values by as much as 500 percent. The results of our simulations are tabulated in Table 2. In these simulations it was assumed that no faulty parameter exceeded a value of 1000 and hence the search was stopped if the minimizing value for $J_i(r_i)$ reached 1000. This was necessitated by the requirement that the Golden Section search be restricted to a finite interval. Of course, the minimization algorithm can be easily modified to take into account infinite values of r_i; i.e. open or short circuited components.

Multiple Failures

Although we have not given any numerical examples of the

application of the search algorithm to the case where multiple failures are assumed, the basic concept of our algorithm remains valid. Computationally, the simple one-dimensional Golden Section search used to minimize $J_i(r_i)$, however, must be replaced by a multidimensional optimization algorithm; say steepest decent, conjugate gradient, etc; to minimize $J_{i(1), i(2), \ldots, i(q)}$ $(r_{i(1)}, r_{i(2)}, \ldots, r_{i(q)})$ for each set of q allowable failures. In addition, each evaluation of $F_{i(1), i(2), \ldots, i(q)}$ requires the inversion of a pxp matrix ($p \geq q$) rather than the simple scalar operations required in the single fault case.

An alternative approach to the minimization of $J_{i(1), i(2), \ldots, i(q)}$ which requires only scalar mathematics is to use a cyclical one-dimensional search algorithm on the q parameters $r_{i(1)}, r_{i(2)}, \ldots, r_{i(q)}$. Here, one minimizes $J_{i(1), i(2), \ldots, i(q)}$ as a function of one parameter at a time cycling through the q parameters until a minimum is achieved. Although such an optimization algorithm usually requires more iterations than a true multidimensional optimization, the entire process can be carried out with scalar operations. In particular, since one only varies a single parameter at a time, Householder's formula allows $F_{i(1), i(2), \ldots i(q)}$ to be evaluated at each interation without matrix inversion while a simple Golden Section search may be used for each one-dimensional minimization.

Robustness

Unlike the case of fault diagnosis in a digital system wherein

a component is unambiguously good or bad,in an analog circuit or system, a component parameter is either in tolerence or out of tolerence. As such, any fault diagnosis algorithm which makes use of the nominal component parameters must be tested for robustness. I.e. how effective is the algorithm at locating the faulty component(s) when the good components are not precisely equal to their nominal values. As such, our search algorithm for fault diagnosis was applied to the transistor amplifier using simulated measurements in which one component was out of tolerence (taken to be 10 percent) and the remaining component parameters were in tolerence but not equal to their nominal values.[7] Of course, the nominal values are used to define the F_i since the actual value of the good components is unknown. Not surprisingly, this results in some ambiguity in the diagnosis process since $J_i(r_i)$ can never be reduced exactly to zero. Fortunately, the data of Table 2 resulting from our "perfect" simulation, indicates that we have five orders of magnitude in which to work. As such, our simulation yielded good though not perfect results. In particular, the algorithm correctly located the fault in 71 percent of the trials with an ambiguity group of one in 50 percent of these cases and ambiguity groups of two, three and four in the remaining cases. Here, the ambiguity group was taken to be the set of all r_i for which the minimum value of J_i was of the smallest order of magnitude achieved by any of the performance measures. The results of some typical simulations are shown in Table 3. Since <u>all</u> of the good components in this simulation were taken to be at the limits of their tolerence interval, these results actually

represent a worst case situation. As such, we believe that the search algorithm will yield significantly better results in a "real world" situation, wherein most of the components will have near nominal values with relatively few of the "good" component parameters lying near their tolerence limits.

Hybrid Algorithms

Although the terminology has only recently been formulated[1], most of the algorithms which have been proposed over the years of the solution of the fault analysis problem in analog circuits and systems can naturally be catagorized as either "simulation before test" or "simulation after test" algorithms.[9] Although the preceding development has been presented in the context of a "simulation after test" algorithm, any of the techniques, such as the application of Householder's formula[5], are also applicable to "simulation before test" algorithms. Indeed, the techniques are ideally suited to a hybrid algorithm. Here, one would employ a two-pass diagnostic algorithm wherein the measured data vector, m, is first compared with pre-simulated data stored in a fault dictionary. If the fault is so located, the diagnosis process is terminated. If the fault is not located among those which have been presimulated and stored in the fault dictionary, the hybrid algorithm will then revert to a "simulation after test" mode until a sequence of parameter vectors, r_i, and simulated data vectors, m_i, have been computed which converge to the solution of the fault diagnosis equations. At the same time the results of each of these "after test" simulations are stored in the fault dictionary

for use in future applications of the test algorithm. As such, a fault dictionary is slowly built up which includes simulations of those failures which are most commonly encountered in actual practice. Such, a hybrid algorithm would seem to achieve the best of both worlds. Common faults would be found quickly on the first pass, yet the system would still have the "simulation after test" algorithm upon which to fall back when encountering a new failure mode. Moreover, ATPG requirements would be greatly reduced with only the most common faults (say open and short circuits, single failures, etc.), being pre-simulated and the remainder of the fault dictionary being adaptively generated by the "simulation after test" algorithm as new fault modes are encountered. Such a hybrid scheme alleviates the necessity of determining the fault modes of a system in advance, as required for "simulation before test" while simultaneously eliminating the duplicate simulations of common faults required for "simulation after test".

Conclusions

Our purpose in the preceding has been the formulation of a class of techniques which we believe can serve as the basis of an effective algorithm for fault diagnosis in linear analog circuits and systems. These techniques have proven to be effective in the situation where all good component parameters are "near" nominal and give promise of sufficient robustness to cope with the "real world" situation, in which the good component parameters are in tolerence though not nominal.

Although the presentation has been formulated in the context of a "simulation after test" algorithm, the techniques presented are also applicable to "simulation before test" and hybrid algorithms.

References

1. "Report of the Industry-Joint Services Automatic Test Task Force", San Diego, April, 1977, (to appear).

2. Saeks, R., Singh, S.P., and R.W. Liu, "Fault Isolation via Components Simulation", IEEE Trans. on Circuit Theory, Vol. CT-19, pp. 634-640, (1972).

3. Ransom, M.N., and R. Saeks, "A Functional Approach to Fault Analysis in Linear Systems", in Rational Fault Analysis, (eds. R. Saeks and S.R. Liberty), New York, Marcel Dekker Inc., 1977, pp. 124-134.

4. Householder, A.S., "A Survey of Some Closed Methods for Inverting Matrices", SIAM Jour. on Appl. Math., Vol. 5, pp. 155-169, (1957).

5. Temes, G.C., "Efficient Methods for Fault Simulation", Proc. of the 20th Midwest Symp. on Circuits and Systems, Texas Tech Univ., Aug. 1977, pp. 191-194.

6. Leung, K.H., and R. Spence, "Multiparameter Large-Change Sensitivity Analysis and Systematic Exploration", IEEE Trans. on Circuits and Systems, Vol. CAS-22, pp. 796-804, (1975).

7. Chen, H.S.M., M.S. Thesis, Texas Tech University, Lubbock, Texas, 1977.

8. Saeks, R., and R.S. DeCarlo, Interconnected Dynamical Systems, New York, Marcel Dekker Inc., 1979 (to appear).

9. Saeks, R., and S.R. Liberty, Rational Fault Analysis, New York, Marcel Dekker Inc., 1977.

10. Sen, N., M.S. Thesis, Texas Tech University, Lubbock, Tx., 1977.

11. Sen, N., and R. Saeks, "A Measure of Testability and its Application to Test Point Selection - Computation", Proc. of AUTOTESTCON '77, pp. 212-219, Hyannis, Mass., Nov. 1977.

12. Sen, N., and R. Saeks, "A Measure of Testability and its Application to Test Point Selection - Theory", Proc. of the 20th Midwest Symp. on Circuits and Systems, Texas Tech Univ., Lubbock, Texas, August 1977, pp. 576-583.

13. Trauboth, H., and N. Prasad, "MARSYAS - A Software Package for Digital Simulation of Physical Systems", Proc. of the Spg. Joint Comp. Conf., pp. 223-235, 1970.

Component	C_1	L_1	C_2	L_2
Actual Parameter Values	20	20	30	40
Minimum Value for J_i	1.25×10^{-12}	5.88×10^{-8}	8.46×10^{-9}	1.96×10^{-6}
Minimizing Component Value	20	30.66	39.87	51.89

Component	C_1	L_1	C_2	L_2
Actual Parameter Values	10	40	30	40
Minimum Value for J_i	1.62×10^{-7}	2.10×10^{-4}	2.42×10^{-7}	6.94×10^{-6}
Minimizing Component Value	28.75	40	48.5	62.2

Component	C_1	L_1	C_2	L_2
Actual Parameter Value	10	20	50	40
Minimum Value for J_i	2.91×10^{-8}	2.73×10^{-7}	1.47×10^{-13}	6.87×10^{-6}
Minimizing Component Value	30.27	41.62	50	64.01

Component	C_1	L_1	C_2	L_2
Actual Parameter Value	10	20	30	45
Minimum Value for J_i	3.93×10^{-7}	4.93×10^{-7}	4.13×10^{-7}	1.04×10^{-13}
Minimizing Component Value	14.18	24.46	34.13	45

Table 1. Fault Analysis for LC Filter

Component	C_1	r_x	r_π	C_u	C_2	R_s	R_e	C_π	C_e	g_m	R_c	R_L
Actual Parameter Value	50	10	40	25	20	75	30	15	10	10	10	20
Minimum Value for J_i	6.2×10^{-14}	2.6×10^{-4}	3.1×10^{-5}	3.0×10^{-5}	1.58×10^{-5}	4.5×10^{-4}	5.3×10^{-5}	3.0×10^{-4}	1.7×10^{-4}	3.4×10^{-4}	1.2×10^{-5}	1.8×10^{-6}
Minimizing Component Value	50	88.5	0.44	28.77	1000	1000	62.6	80.1	9.2	4.5	18.2	41.9

Component	C_1	r_x	r_π	C_u	C_2	R_s	R_e	C_π	C_e	g_m	R_c	R_L
Actual Parameter Value	20	30	40	25	20	75	30	15	10	10	10	20
Minimum Value for J_i	1.1×10^{-3}	3.2×10^{-12}	2.1×10^{-2}	2.1×10^{-3}	1.3×10^{-3}	1.3×10^{-3}	8.5×10^{-3}	1.9×10^{-2}	3.5×10^{-4}	1.6×10^{-2}	7.4×10^{-4}	9.0×10^{-4}
Minimizing Component Value	1.68	30	1000	14.6	0.9	0.96	7.56	0.1	16.9	1000	1.2	2.0

Component	C_1	r_x	r_π	C_u	C_2	R_s	R_e	C_π	C_e	g_m	R_c	R_L
Actual Parameter Value	20	10	90	25	20	75	30	15	10	10	10	20
Minimum Value for J_i	1.1×10^{-9}	7.8×10^{-9}	7.6×10^{-20}	3.3×10^{-11}	2.1×10^{-9}	2.2×10^{-9}	5.2×10^{-9}	1.2×10^{-8}	4.1×10^{-9}	1.3×10^{-8}	6.2×10^{-10}	6.9×10^{-10}
Minimizing Component Value	19.9	10.01	90	24.97	19.8	72.7	29.9	14.6	10.01	10.09	9.96	19.9

Table 2. Fault Analysis for Transistor Amplifier

Component	C_1	r_x	r_π	C_u	C_2	R_s	R_e	C_π	C_e	g_m	R_c	R_L
Actual Parameter Value	20	10	40	60	20	75	30	15	10	10	10	20
Minimum Value for J_i	5.6×10^{-3}	3.4×10^{-3}	3.0×10^{-3}	8.1×10^{-13}	9.5×10^{-3}	1.4×10^{-2}	6.0×10^{-3}	1.7×10^{-3}	2.0×10^{-3}	3.1×10^{-3}	2.7×10^{-3}	4.2×10^{-3}
Minimizing Component Value	1000	5.79	0.1	60	1000	1000	1000	1000	6.65	0.72	1000	1000

Component	C_1	r_x	r_π	C_u	C_2	R_s	R_e	C_π	C_e	g_m	R_c	R_L
Actual Parameter Value	20	10	40	25	50	75	30	15	10	10	10	20
Minimum Value for J_i	4×10^{-9}	9.7×10^{-5}	1.6×10^{-5}	1.7×10^{-5}	1.7×10^{-14}	8×10^{-5}	6.3×10^{-6}	1.2×10^{-4}	6.8×10^{-5}	1.2×10^{-4}	1.4×10^{-5}	6.9×10^{-6}
Minimizing Component Value	30.9	9.4	0.88	26.9	50	1000	43	43	9.6	6.3	13.5	28.7

Component	C_1	r_x	r_π	C_u	C_2	R_s	R_e	C_π	C_e	g_m	R_c	R_L
Actual Parameter Value	20	10	40	25	20	15	30	15	10	10	10	20
Minimum Value for J_i	5.2×10^{-6}	1.6×10^{-4}	4.7×10^{-4}	2.1×10^{-4}	4.1×10^{-8}	9.1×10^{-13}	3.2×10^{-5}	3.1×10^{-4}	1×10^{-4}	2.4×10^{-4}	1.8×10^{-5}	9.4×10^{-6}
Minimizing Component Value	12.3	11.1	1000	22.6	9.5	15	20.5	0.1	10.8	50.9	6.9	13

Table 2. Fault Analysis for Transistor Amplifier (Cont:)

Component	C_1	r_x	r_π	C_u	C_2	R_s	R_e	C_π	C_e	g_m	R_c	R_L
Actual Parameter Value	20	10	40	25	20	75	80	15	10	10	10	20
Minimum Value for J_i	6.8×10^{-5}	5.5×10^{-4}	1.5×10^{-4}	1.5×10^{-4}	4.3×10^{-5}	6.6×10^{-4}	2.2×10^{-14}	5.9×10^{-4}	4.2×10^{-4}	6.4×10^{-4}	1.4×10^{-4}	8.8×10^{-5}
Minimizing Component Value	61.4	8.9	0.4	29.2	1000	1000	80	78.3	9.2	4.6	19.6	47.8

Component	C_1	r_x	r_π	C_u	C_2	R_s	R_e	C_π	C_e	g_m	R_c	R_L
Actual Parameter Value	20	10	40	25	20	75	30	55	10	10	10	20
Minimum Value For J_i	8.5×10^{-5}	2.5×10^{-6}	5.4×10^{-5}	5.2×10^{-5}	1.1×10^{-4}	1.3×10^{-4}	1.3×10^{-4}	3.5×10^{-14}	1×10^{-5}	9.6×10^{-7}	6.3×10^{-5}	7.6×10^{-5}
Minimizing Component Value	27.4	9.2	1	26.7	34.6	1000	37.1	55	9.5	5.3	12.8	26.4

Component	C_1	r_x	r_π	C_u	C_2	R_s	R_e	C_π	C_e	g_m	R_c	R_L
Actual Parameter Value	20	10	40	25	20	75	30	15	30	10	10	20
Minimum Value for J_i	2.6×10^{-3}	2.4×10^{-3}	7.6×10^{-2}	6.1×10^{-4}	3.1×10^{-3}	3.1×10^{-3}	1.3×10^{-2}	7.3×10^{-2}	1.9×10^{-11}	6.7×10^{-2}	2.9×10^{-3}	2.5×10^{-3}
Minimizing Component Value	0.1	1000	1000	10	0.1	0.1	2.7	0.1	30	1000	0.1	0.1

Table 2. Fault Analysis for Transistor Amplifier (Cont:)

Component	C_1	r_x	r_π	C	C_2	R_s	R_e	C_π	C_e	g_m	R_c	R_L
Actual Parameter Value	20	10	40	25	20	75	30	15	10	40	10	20
Minimum Value for J_i	8.9×10^{-5}	9.4×10^{-6}	2×10^{-4}	6.2×10^{-5}	1.1×10^{-4}	1.1×10^{-4}	1.3×10^{-4}	5.7×10^{-5}	2×10^{-5}	5.7×10^{-15}	7×10^{-5}	8×10^{-5}
Minimizing Component Value	15.4	10.9	1000	23.6	13.5	27.6	25.4	0.1	10.5	40	8.1	15.8

Component	C_1	r_x	r_π	C_u	C_2	R_s	R_e	C_π	C_e	g_m	R_c	R_L
Actual Parameter Value	20	10	40	25	20	75	30	15	10	10	50	20
Minimum Value for J_i	6.9×10^{-5}	8.3×10^{-4}	6.7×10^{-5}	5.3×10^{-5}	8.4×10^{-4}	2.3×10^{-3}	3.4×10^{-4}	2.3×10^{-3}	5.1×10^{-4}	1.1×10^{-3}	4.3×10^{-14}	6.2×10^{-5}
Minimizing Component Value	1000	7.8	0.2	33.9	1000	1000	1000	1000	8.4	2.4	50	1000

Component	C_1	r_x	r_π	C_u	C_2	R_s	R_e	C_π	C_e	g_m	R_c	R_L
Actual Parameter Value	20	10	40	25	20	75	30	15	10	10	10	60
Minimum Value for J_i	3.3×10^{-6}	4.2×10^{-4}	4.1×10^{-5}	3.7×10^{-5}	1.3×10^{-4}	9.1×10^{-4}	1.1×10^{-4}	3.9×10^{-4}	2.7×10^{-4}	5.6×10^{-4}	9.1×10^{-6}	4.7×10^{-14}
Minimizing Component Value	84	8.5	0.3	30.3	1000	1000	93.9	1000	8.9	3.6	23.7	60

Table 2. Fault Analysis for Transistor Amplifier (Cont:).

Component	C_1	r_x	r_π	C_u	C_2	R_s	R_e	C_π	C_e	g_m	R_c	R_L
Nominal Parameter Value	20	10	40	25	20	75	30	15	10	10	10	20
Actual Parameter Values	22	11	36	62	22	82	33	14	9	11	9	22
Minimum Value for J_i	7.5×10^{-3}	4.8×10^{-3}	4.7×10^{-3}	2×10^{-5}	1.2×10^{-2}	1.7×10^{-2}	7.9×10^{-3}	3×10^{-3}	2.9×10^{-3}	4.1×10^{-3}	4.2×10^{-3}	6×10^{-3}
Minimizing Component Value	1000	5.6	0.1	69.9	1000	1000	1000	1000	6.4	0.6	1000	1000
Ambiguity Set(X) Actual Fault(0)				X								

Component	C_1	r_x	r_π	C_u	C_2	R_s	R_e	C_π	C_e	g_m	R_c	R_L
Nominal Parameter Values	20	10	40	25	20	75	30	15	10	10	10	20
Actual Parameter Values	22	11	36	27	22	2	33	14	9	11	9	22
Minimum Value for J_i	2.6×10^{-5}	1.3×10^{-3}	7.6×10^{-3}	1.6×10^{-4}	2.6×10^{-6}	2.5×10^{-6}	1.2×10^{-3}	6.7×10^{-3}	6.5×10^{-4}	5.3×10^{-3}	1.1×10^{-4}	5.7×10^{-5}
Minimizing Component Value	4.2	17	1000	17.4	2.52	2.9	10.1	0.1	13.7	1000	2.8	4.9
Ambiguity Set (X) Actual Fault (0)					X	X						

Table 3. Fault Analysis of Transistor Amplifier with 10 percent tolerences

Component	C_1	r_x	r_π	C_u	C_2	R_s	R_e	C_π	C_e	g_m	R_c	R_L
Nominal Parameter Values	20	10	40	25	20	75	30	15	10	10	10	20
Actual Parameter Value	22	11	36	27	22	82	33	37	9	11	9	22
Minimum Value for J_i	7.9×10^{-5}	2.6×10^{-4}	1.4×10^{-6}	1.1×10^{-6}	2.4×10^{-4}	1×10^{-3}	2.2×10^{-4}	2.7×10^{-4}	1.3×10^{-4}	3.6×10^{-4}	3.3×10^{-5}	5.3×10^{-5}
Minimizing Component Value	86.5	8.3	0.3	30.7	1000	1000	94.7	140	8.8	3.1	24.6	62
Ambiguity Set (X) Actual Fault (0)			X	X				0				

Table 3. Fault Analysis of Transistor Amplifier with 10 percent tolerences (Cont:).

RESEARCH

on

COMPUTER-AIDED DESIGN

C.T. Pan and K.S. Chao

DEPARTMENT OF ELECTRICAL ENGINEERING

TEXAS TECH UNIVERSITY

ABSTRACT

An efficient computer-aided root-locus method is described. The approach is based on the concept of continuation methods in which the solution of a parameterized family of algebraic problems is converted into the solution of a differential equation. The root-locus plot is obtained in a systematic manner by numerical integration. Singularities are analyzed and classified according to the properties of higher order derivaties. Depdending on their classification, singular points on the root loci are taken care of accordingly.

INTRODUCTION

The root-locus method is one of the most important design techniques for linear time-invariant feedback systems. In addition to yielding frequency response information of the system, it also provides a powerful tool for solving problems in the time domain. The basic idea of the root-locus method is to determine the closed-loop pole configuration as a function of the gain from the configuration of the open-loop poles and zeros. A great deal of information is available in texts and literature on the method for the construction of root loci. The graphical method using certain elementary geometric properties of the locus is probably the most commonly used approach (see e.g. [1] - [3]). Other approach [4] - [7] employ either analytic or semi-analytic representations that involve the use of equations of the loci. Although analytic approaches enable one to obtain accurate plots along with certain qualitative features of the root paths, the point-to-point plotting is just a formidable task. Besides, investiations for higher order systems are virtually impractical.

It is the purpose of this paper to develop a computer-aided method for plotting root loci in a systematic manner. The approach to be presented in Section II is based on the concept of continuation methods [8] - [10]. The basic idea is to convert the solution of a parameterized family of algebraic problems into the solution of a set of associated differential equations. Section III is concerned with the existence and classification of singular points on the root loci. In section IV the results obtained are illustrated by means of examples.

The Root-Locus Method

Consider the closed-loop system shown in Fig. 1. Let the open-loop transfer function be expressed by

$$G(s)H(s) = K\frac{A(s)}{B(s)} \triangleq K\frac{(s-z_1)(s-z_2)\ldots(s-z_m)}{(s-p_1)(s-p_2)\ldots(s-p_n)} \tag{1}$$

where K is the open-loop gain and $m < n$. The closed-loop transfer function is

$$T(s) = \frac{G(s)}{1 + G(s)H(s)} = \frac{G(s)\ B(s)}{B(s) + KA(s)} . \tag{2}$$

The root-locus plot of the closed-loop transfer function T(s) is defined as the locus of the poles of T(s) when K varies from zero to infinity. This plot consists of a set, denoted by ℓ, of points in the s-plane such that

$$g(s,K) \triangleq B(s) + KA(s) = 0, \tag{3}$$

i.e.,

$$\ell = \{s \mid g(s,K) = 0\}. \tag{4}$$

Instead of solving the roots of (3) directly for each K, a system of two simultaneous differential equations

$$\frac{d}{dt}\, g(s(t),\, K(t)) = -g(s(t),\, K(t)), \quad g(s(0),\, K(0)) = 0$$

$$\frac{d}{dt}\, K(t) = \pm\, 1, \qquad K(0) = K_o \tag{5}$$

is considered where $S(0) = S_o$ is a root of [3] corresponding to an initial gain K_o, and t is a dummy variable. Application of the chain-rule to (5) results in

$$\frac{ds}{dt} = -(g + \frac{\partial g}{\partial K}) / \frac{\partial g}{\partial s}\,, \quad s(0) = s_o$$

$$\frac{dK}{dt} = \pm\, 1 \tag{6}$$

or equivalently,

$$\frac{ds}{dt} = -\frac{(B(s) + KA(s)) + A(s)}{B'(s) + KA'(s)}\,, \quad s(0) = s_o$$

$$\frac{dK}{dt} = \pm\, 1. \qquad K(0) = K_o \tag{7}$$

Equation (7) can now be solved by any numerical integration technique. For example, using Euler's Method, (6) reduces to

$$s_{k+1} = s_k - h\, \frac{B(s_k) + K_k A(s_k) + A(s_k)}{B'(s_k) + K_k A'(s_k)} \tag{8}$$

$$K_{k+1} = K_k \pm h.$$

It is seen from the solution of (5)

$$g(s,K) = g(s(0),K(0))e^{-t} = 0e^{-t} \equiv 0$$

$$K = \pm\, t$$

that for any admissible pair K_o and s_o satisfying (3), the corresponding trajectory resulted from (5) will remain on the solution curve $g(s,K) = 0$ as K changes. The + or - sign is chosen depending on whether one would like to increase or decrease K. Since the computed trajectory may not satisfy (3) exactly, the minus sign in front of g in (5) is used to ensure that the computed trajectory does not diverge away from the locus.

It is a well known fact that the root-locus plot for T(s) contains n branches starting from the open-loop poles at K = 0. Therefore, in the case where the open-loop poles are distinct, the n initial conditions for (7) are selected at $K(0) = 0$ and $s(0) = p_i$, $i = 1,2,\ldots,n$. In the case where the open-loop transfer function contains repeated poles, the term $(\partial g/\partial s)$ becomes zero when evaluated at the repeated poles. As a result, the selection of starting points cannot be made at K = 0 for the repeated poles. Approximate starting points, however, can easily be obtained by analyzing the properties of the root loci in the neighborhood of the repeated pole. Suppose the open loop gain has a repeated pole p_k with multiplicity r and the corresponding g(s,K) has the form

$$g(s,K) = B(s) + KA(s)$$

$$= (s-p_1)(s-p_2)\ldots(s-p_k)^r\ldots(s-p_{n-r}) + KA(s) = 0. \quad (9)$$

Let

$$w = p_k + \Delta s = p_k + \varepsilon e^{j\theta} \tag{10}$$

where ε is an arbitrarily small real number and θ is a phase angle yet to be determined. Thus at $s = w$

$$g(w,K) = g(p_k + \varepsilon e^{j\theta}, K) \simeq 0 \tag{11}$$

Solving Δs from (11), gives

$$\Delta s = \varepsilon e^{j\theta} = (K\rho e^{j\phi})^{1/r} \tag{12}$$

where

$$\rho e^{j\phi} = - \left. \frac{A(s)(s-p_k)^r}{B(s)} \right|_{s=p_k} \tag{13}$$

Therefore r approximate starting points in the neighborhood of the repeated pole p_k and the corresponding open-loop gain can be evaluated from (12) as

$$w_i = p_k + \varepsilon e^{j(\frac{\phi+2\pi i}{r})} \quad , \quad i = 0,1,2,\ldots r-1$$

and

$$K = \frac{1}{\rho} e^r. \tag{14}$$

With the proper choice of n starting points, the n branches of the root-locus plot can be traced in a continuous manner by numerical integration. In computing the root locus, care must be exercised when approaching a singular point on the locus.

Singular Points

A point s* satisfying

$$\left.\frac{dK}{ds}\right|_{s=s^*} = \left.\frac{d}{ds}\left(-\frac{B(s)}{A(s)}\right)\right|_{s=s^*}$$

$$= -\frac{A(s^*)B'(s^*) - B(s^*)A'(s^*)}{A^2(s^*)} = 0 \qquad (15)$$

in the complex plane is called a singular point. Since the numerator of (15) is a (n+m-1)th order polynomial of real coefficients, there are (n+m=1) singular points in the s-plane. Only those singular points that are located on the root loci will be considered. In view of the fact that A(s*) cannot be zero for a finite K, it follows that on the root locus, the condition

$$A(s^*)B'(s^*) - B(s^*)A'(s^*) = A(s^*)(B'(s^*) + KA'(s^*)) = 0 \qquad (16)$$

implies

$$\left.\frac{\partial g}{\partial s}\right|_{s=s^*} = B'(s^*) + KA'(s^*) = 0. \qquad (17)$$

Hence (7) is not valid at s = s* and modifications must be made to handle these singular points. Condition (15) does include the conventional break-in and break-away points at which K is either a local maxima or a local minimum on the real axis, respectively. In general, $\frac{dK}{ds}$ is a complex quantity

if s is not located on the real axis and it does not make sense to talk about local extremal values without proper modification. Now, since K is a real valued function of s for all s on the root locus ℓ, the directional derivative $\frac{dK}{d\ell}$ together with its higher order derivatives along the tangential direction of the locus are well defined. It is thus possible to consider local extremal values of K along ℓ using the notion of directional derivatives.

Let

$$K(s) = \frac{B(s)}{A(s)} \triangleq U(x,y) + jV(x,y) \tag{18}$$

where U and V are real valued functions of x and y, and

$$s = x + jy. \tag{19}$$

The directional derivative of K at a point $s \varepsilon \ell$ in the unit direction $v = e^{j\theta} = v_x + jv_y$ tangential to the root locus is

$$\frac{dK}{d\ell} = \nabla U(x,y) \cdot v = \frac{\partial U}{\partial x} v_x + \frac{\partial U}{\partial y} v_y . \tag{20}$$

The above equation can also be written in the form

$$\frac{dK}{d\ell} = \mathrm{Re}[(\frac{\partial U}{\partial x} - j \frac{\partial U}{\partial y})(v_x + jV_y)], \; s \varepsilon \ell . \tag{21}$$

Making use of the Cauchy-Riemann condition, (21) reduces to

$$\frac{dK}{d\ell} = \mathrm{Re}\,[(\frac{\partial U}{\partial x} + j \frac{\partial V}{\partial x}) e^{j\theta}]$$

$$= \mathrm{Re}[K'(s) e^{j\theta}] \tag{22}$$

where $K' = dK/ds$.

Similarly, higher order directional derivatives of K with respect to ℓ are related to the higher order derivatives by

$$\frac{d^m K}{d\ell^m} = R_e[K^{(m)}(s)e^{jm\theta}]. \tag{23}$$

Thus it is seen from (23) that along ℓ, K and its directional derivatives are all real-valued functions. The following theorem which plays an important role in singularity classification will be proved.

Theorem

Suppose p(s) is an analytic function such that

$$p(s^*) = a, \quad \text{for a real } a \neq 0,$$

$$p^{(k)}(s^*)=0, \quad k = 1,2,\ldots,q-1,$$

and $$p^{(q)}(s^*) \neq 0, 2 \leq q \leq n$$

at some point s^* located on $Imp(s) = 0$. Let

$$R_p = \{s \mid Imp(s) = 0\} .$$

Then in the neighborhood of s^*, R_p consists of q branches, $R_{p1}, R_{p2}, \ldots, R_{pq}$, and $R_{pi} \cap R_{p2} \cap \ldots \cap R_{pq} = s^*$. Furthermore, for each i, $1 \leq i \leq q$, $Rep(s)\big|_{R_{pi}}$ is either a local maximum or a local minimum at s^* if q is even; it is either an increasing function or a decreasing function if q is odd.

Proof:

Without loss of generality s^* can be assumed to be zero. Before considering the general case, the theorem is proved for

$$h(s) = a + s^q.$$

Identifying the set R_h from

$$R_h = \{S | Imh(s) = 0\}$$

yields

$$R_h = \{re^{j\theta} | r^q \sin q\theta = 0\}$$

$$= \{re^{j\theta} | \theta = i\pi/q,\ i = 0,1,2,\ldots q-1\}$$

where θ is restricted in the upper half of the s-plane and r assumes negative values in the lower half plane. Thus R_h consists of q intersecting branches R_{hi}, $i = 0,1,2,\ldots,$ $q-1$. The intersection occurs at $r = 0$. For each i,

$$Reh(s)\Big|_{R_{hi}} = a + r^q\cos(q\theta_i) = a + r^q\cos(i\pi).$$

Therefore if q is even, r^q is always nonnegative, and

$$Reh(s)\Big|_{R_{hi}} \begin{cases} \geq a & \text{when } \cos i\pi = 1 \\ \leq a & \text{when } \cos i\pi = -1, \end{cases}$$

i.e., $h(s)\Big|_{R_{hi}}$ is either a local maximum or a local minimum. Now if q is odd, then for each i,

$$\mathrm{Re}\,h(s)\Big|_{R_{hi}} = a + r^q \cos(i\pi).$$

Hence $(h(s)\Big|_{R_{hi}} - a)$ will change sign either from plus to minus as r increases from negative value to positive value or vice versa, i.e., $h(s)\Big|_{R_{hi}}$ is a monotonic function.

Returning to the general case, p(s) can be expanded into a Taylor series around s* as

$$p(s) = p(s^*) + \sum_{i=q}^{\infty} c_i (s-s^*)^i$$

$$= a + (s-s^*)^q \sum_{k=0}^{\infty} c_{q+k} (s-s^*)^k.$$

Since the summation in the above equation is an analytic function and has no zero in a small disc around s*, from a theorem in the theory of complex variables (see e.g. [11]), there exists an analytic function u(s) such that

$$\sum_{k=0}^{\infty} c_{k+q} (s-s^*)^k = \exp(u(s)).$$

Let $v(s) = \exp\,(u(s)/q)$. Then

$$p(s) = a + [(s-s^*)v(s)]^q \triangleq a + [f(s)]^q$$

where $f(s^*) = 0$, $f'(s^*) \neq 0$. Thus f(s) is a local homeomorphism. Now

$$R_p = \{s \mid \mathrm{Im}(a + (f(s))^q) = 0\}$$

$$= \{s \mid f(s) \in R_h\} \quad .$$

Let

$$R_{pi} = \{s \mid f(x) \in R_{hi}\}\ ,\ i = 0,1,2,\ldots,q-1.$$

If q is even, then $s \in R_{pi}$ implies $f(s) \in R_{hi}$. It follows that

$$\mathrm{Reh}(f(s)) \begin{matrix} \geq \\ \leq \end{matrix} \mathrm{Reh}(0) = \mathrm{Reh}(f(s^*)), \quad \begin{matrix} \text{when } \cos i\pi = 1 \\ \text{when } \cos i\pi = -1\ , \end{matrix}$$

i.e.,

$$\mathrm{Rep}(s)\Big|_{R_{pi}} \begin{matrix} \geq \\ \leq \end{matrix} \mathrm{Rep}(s^*), \quad \begin{matrix} \text{when } \cos i\pi = 1 \\ \text{when } \cos i\pi = -1. \end{matrix}$$

Therefore $\mathrm{Rep}(s)\Big|_{R_{pi}}$ is either a local maximum or a local minimum at s^*. Similarly if q is odd, then it can be shown that $\mathrm{Rep}(s)\Big|_{R_{pi}}$ is either an increasing function or a decreasing function.

As a directly consequence of the above theorem, the following corollary is deduced.

Corollary

Suppose s^* is a singular point on ℓ such that

$$K(s^*) \neq 0$$

$$\frac{d^k K}{ds^k}\Big|_{s=s^*} = 0, \quad k = 1,2,\ldots,q-1$$

$$\frac{d^q K}{ds^q}\Big|_{s=s^*} \neq 0, \quad 2 \leq q \leq n$$

where $K(s) = -B(s)/A(s)$, then there are q branches intersecting at $s = s^*$. Furthermore if q is even, then along each branch of the intersecting root loci at $s = s^*$, $K(s^*)$ is either a local maximum or a local minimum; otherwise $K(s^*)$ is a monotonic function of s on that branch in the neighborhood of s^*.

Proof:

Let $f(s) = - B(s)/A(s)$. Then f(s) is an analytic function. It follows that

$$B(s) + f(s)A(s) = 0.$$

Comparing the above equation to (3), it is obvious that

$$\begin{aligned} K &= \mathrm{Re}f(s) \\ 0 &= \mathrm{Im}f(s) \end{aligned} \qquad \text{for all } s \in \ell.$$

Application of the above theorem to f(s) completes the proof.

According to the above corollary, singular points are characterized by the properties of higher order derivatives. It is noted that

$$\frac{d^k K}{ds^k} = \frac{d^k}{ds^k}\,[-B(s)/A(s)] = -\frac{\partial^k q}{\partial s^k}\,/A(s). \qquad (24)$$

Since g(s) is an nth-order polynomial with real coefficients, $\partial^k g/\partial s^k$ can easily be generated. Furthermore, q can at most be equal to n since

$$\frac{\partial^n g(s)}{\partial s^n} = - \frac{n!}{A(s)} \neq 0 \qquad (25)$$

With the above corollary, the conventional break-in and break-away points defined on the real axis can now be generalized as follows.

Definition

In the theorem, the singular point is called an even singular point if q is even; otherwise, it is said to be an odd singular point.

It is clear from the corollary that on the root locus an even singular point is either a local maximum or a local mininum along the branch as defined in the Theorem. The conventional break-in and break-away points are just special cases of even singular points of order 2(i.e., q = 2) which are located on the real axis. In general, if a singular point is located off the real axis, the concept of directional derivatives can be used to characterize the singular point. From (23) and the corollary it is obvious that at an qth-order singular point, $d^k K/d^k \ell = 0$ for k = 1,2,...,q-1 and $d^q K/d\ell^q \neq 0$ There are q branches of the root loci intersecting at the singular points.

In generating the root-locus plot, each branch is plotted separately as K increases. In the neighborhood of an odd singular point, since K is a monotonic function of s on the locus, the first order directional derivative is a continuous function. Thus when approaching an odd singular point, it is

necessary to jump over the singular point by adding a small variation $|\Delta s|$ along the tangential direction of the locus. For even singular points, such procedure is invalid. Since on the root-locus plot an even singular point is either a local maximum or a local minimum, such construction does not give rise to an increasing K. Thus in order to cintinue the plotting of the root locus as a function of increasing K, it is necessary to change the direction of the locus when an even singular point is approached. Depending on the order q of the even singular point, Δz, the change in direction, is chosen as

$$\Delta z = \Delta s e^{-j\pi/q} \tag{26}$$

where Δs is a sufficiently small vector in the tangential direction of the locus when approaching the singular point. The factor $e^{-j\pi/q}$ can be viewed as a rotational operator, which rotates the direction clockwise by π/g. On a branch so constructed, the even singular point no longer has the characteristics of a local extremum.

It is apparent from the foregoing root-locus construction that the q branches will directly intersect each other at an odd singular point and that the modified q root loci will touch each other at an even singular point. For obvious reasons, an odd singular point is known as an intersecting point while an even singular point is called a touching point. The graphical illustrations for these two types of singularities are shown

in Fig. 2 and Fig. 3 for q = 2 and q = 3, respectively. The branches are numbered and the arrows are pointed in the direction of increasing K.

After the change of direction at a touching point or the jump over an intersecting point, it may be necessary to make corrections if the point selected is not close enough to the locus. This can usually be achieved within a few steps by using the Newton iteration

$$s_{k+1} = s_k - \frac{B(s_k) + KA(s_k)}{B'(s_k) + KA'(s_k)} . \tag{27}$$

The gain K which corresponds to the corrected point can be evaluated either from

$$K = -ReB(s)/ReA(s) \tag{28}$$

or

$$K = -ImB(s)/ImA(s). \tag{29}$$

Equations (28) and (29) are derived by taking the real and the imaginary parts of g(s) = 0, respectively.

Examples

In this section a number of examples are presented to illustrate the proposed algorithm for obtaining the root-locus plot. Although any integration technique can be used to solve (7), only the Euler's method with variable step size is used for illustration.

Example 1:

Consider a linear feedback system whose open loop transfer function is given by

$$G(s)H(s) = K\frac{(s + 1.5)(s + 5.5)}{s(s + 1)(s + 5)} \quad .$$

there are three simple poles at 0, -1 and -5, and two finite zeros at -1.5 and -5.5. Application of (7) with starting points 0, -1, and -5 at K = 0 leads to three root loci shown in Fig. 4. It turns out that all 4 singular points are located on the root-locus plot and they are all classified as even singular points with q = 2. It is thus necessary to change the direction of the root locus when each singular point is approached. For branch 1, when Q_1 is approached, the tangential direction $\Delta s = -\varepsilon$ and Δz_1 is chosen as $(-\varepsilon)e^{j\pi/2} = -\varepsilon j$. Similarly, when branch 1 approaches Q_2, Q_3 and Q_4, the changes in direction are chosen as $\Delta z_2 = (j\varepsilon)e^{j\pi/2}$, $\Delta z_3 = (-\varepsilon)e^{j\pi/2}$, and $\Delta z_4 = (j\varepsilon)e^{j\pi/2}$, respectively. Other branches, denoted by 2 and 3, are obtained in a similar manner.

Example 2:

As an example of the case with multiple poles consider

$$G(s)H(s) = \frac{K}{(s + 3)(s + 1)^2}$$

where s = -1 is a repeated pole with multiplicity r = 2. From (14)

$$w_o = 1+\varepsilon e^{j\phi/2}$$

$$w_1 = 1+\varepsilon e^{j(\phi+2\pi)/2}$$

$$K = \frac{1}{\rho}\,\varepsilon^2$$

where

$$\rho e^{j\phi} = -\left.\frac{1}{s+3}\right|_{s=-1} = 0.5e^{j\pi}$$

and ε is chosen as 0.2 for illustration. Thus the two approximate starting points are

$$w_o = -1 + j0.2, \qquad K = 0.08$$

$$w_1 = -1 - j0.2, \qquad K = 0.08.$$

The root loci are obtained by using (8) with $s(0) = w_o, w_1, -3$ and $K(0) = 0.08, 0.08, 0$, respectively. The results are shown in Fig. 5 where the root loci are plotted up to $K = 5$.

Example 3:

In this example, the open-loop transfer function is assumed to have one real pole, and two complex conjugate poles:

$$G(s)H(s) = \frac{K}{s(s+3+j\sqrt{3})(s+3-j\sqrt{3})} .$$

There is only one odd singular point with $q = 3$ located on the root loci. It is thus necessary to jump over the singular point when it is approached and Δz is chosen in the tangential

direction of the locus. The complete root-locus plot is shown in Fig. 6.

Example 4:

The final example demonstrates the case where the singular points are located off the real axis. Consider

$$G(s)H(s) = K\frac{A(s)}{B(s)} = \frac{K}{(s+1)^2(s+1+j\sqrt{18})(s+1-j\sqrt{18})}.$$

It is seen that

$$g(s) = B(s) + KA(s) = s^4 + 4s^3 + 24s^2 + 40s + 19 + K.$$

Setting the derivative of g with respect to s to zero, yields three singular points, namely, $s_1 = -1$, $s_2 = -1+j3$ and $s_3 = -1-j3$. Since $s_1 = -1$ is a repeated open loop pole, it can be taken care of as a starting point. At s_2 and s_3, it is easily verified that

$$\text{Img}(s_2) = \text{Img}(s_3) = 0$$

and

$$\left.\frac{\partial^2 g}{\partial s^2}\right|_{s=s_1 \text{ or } s_2} \neq 0 .$$

This indicates that both s_2 and s_3 are located on the root loci and, furthermore, they are classified as even singular points with q = 2. Application of the proposed root-locus

plotting procedure with four starting points

$$s_{10} = -1 + j\varepsilon, \quad K = 0.18 \quad , \quad \varepsilon = 0.1$$

$$s_{20} = -1 - j\varepsilon, \quad K = 0.18 \quad , \quad \varepsilon = 0.1$$

$$s_{30} = -1 + j\sqrt{18}, \quad K = 0$$

$$s_{40} = -1 - j\sqrt{18}, \quad K = 0$$

leads to the complete root-locus plot shown in Fig. 7. It is clear from this example that a necessary condition for the existence of complex singular points is that the order of the open-loop transfer function be greater than or equal to 4.

Conclusion

An algorithm for generating the root-locus plot has been presented. Classification of singular points has also been discussed in detail. It is shown that the conventional break-in and break-away points are just special cases of even singular points. The computer-aided method successfully solves the problem of discontinuity of the direction of the locus at singular points and enables one to plot the root loci without missing or repeating any branch.

References

[1] W. R. Evans, Control-System Dynamics, New York: McGraw-Hill, 1954.

[2] H. Raven, Automatic Control Engineering, New York: McGraw-Hill, 1961.

[3] C. Gupta, L. Hasdorff, Fundamentals of Automatic Control, New York: Wiley, 1970.

[4] V. Krishnan, "Semi-analytic approach to root locus," IEEE Trans. on Automatic Control, Vol. AC-11, No. 1, pp. 102-108, Jan. 1966.

[5] K. Steiglitz, "An analytic approach to root loci," IRE Trans. on Automatic Control, Vol. AC-6, pp. 326-332, Sept. 1961.

[6] C. K. Wojeik, "Analytical representation of root locus," J. Basic Eng., pp. 37-43, March 1964.

[7] B. P. Bhattachryya, "Root locus equations of the fourth degree," Internat'l J. of Control, Vol. 1, No. 6 pp. 533-556, 1965.

[8] F. H. Branin, "Widely covergent method for finding multiple solutions of simultaneous nonlinear equations," IBM J. Res. Develop., Vol. 16, No. 5, pp. 504-522, Sept. 1972.

[9] K.S. Chao, D. K. Liu and C. T. Pan, "A systematic search method for obtaining multiple solutions of simultaneous nonlinear equations," IEEE Trans. Circuits Syst., Vol. CAS-22, No. 9, pp. 748-753, Sept. 1975.

[10] K.S. Chao and R. Saeks, "Continuation methods in circuit analysis," Proc. IEEE, Vol. 65, No. 8, pp. 1187-1194, Aug. 1977.

[11] W. Rudin, Real and Complex Analysis, New York: McGraw-Hill, 1966.

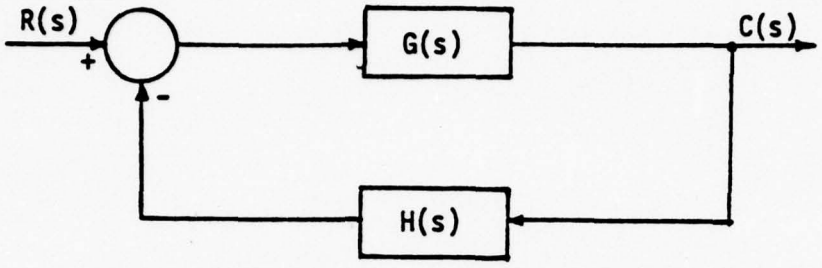


Fig. 1. A closed-loop feedback system.

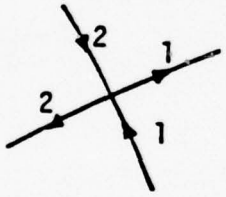


Fig. 2. An even sigular point.

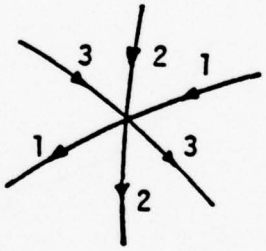


Fig. 3. An odd singular point.

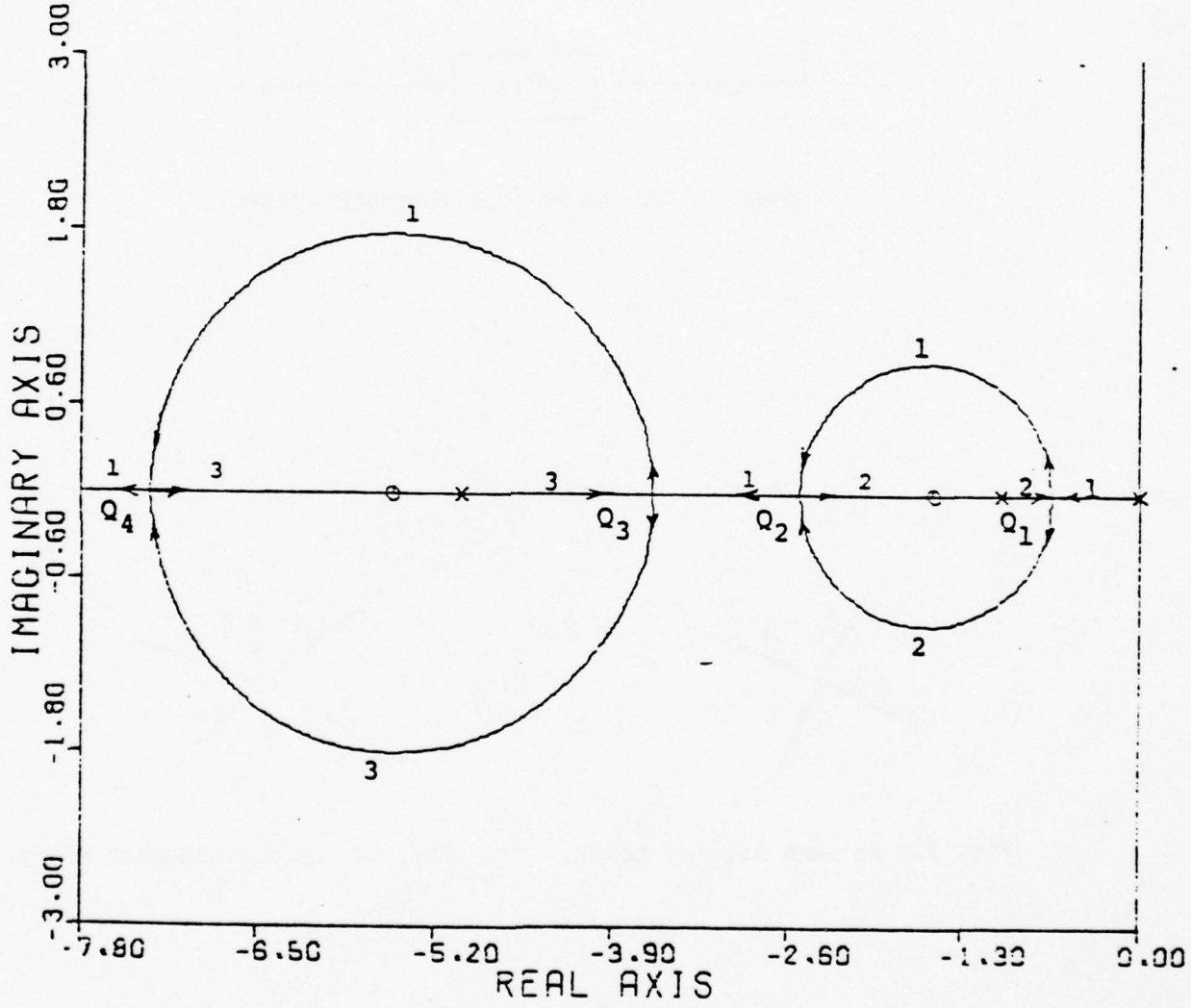


Fig. 4. $G(s)H(s) = \frac{s + 1.5)(s + 5.5)}{s(s + 1)(s + 5)}$.

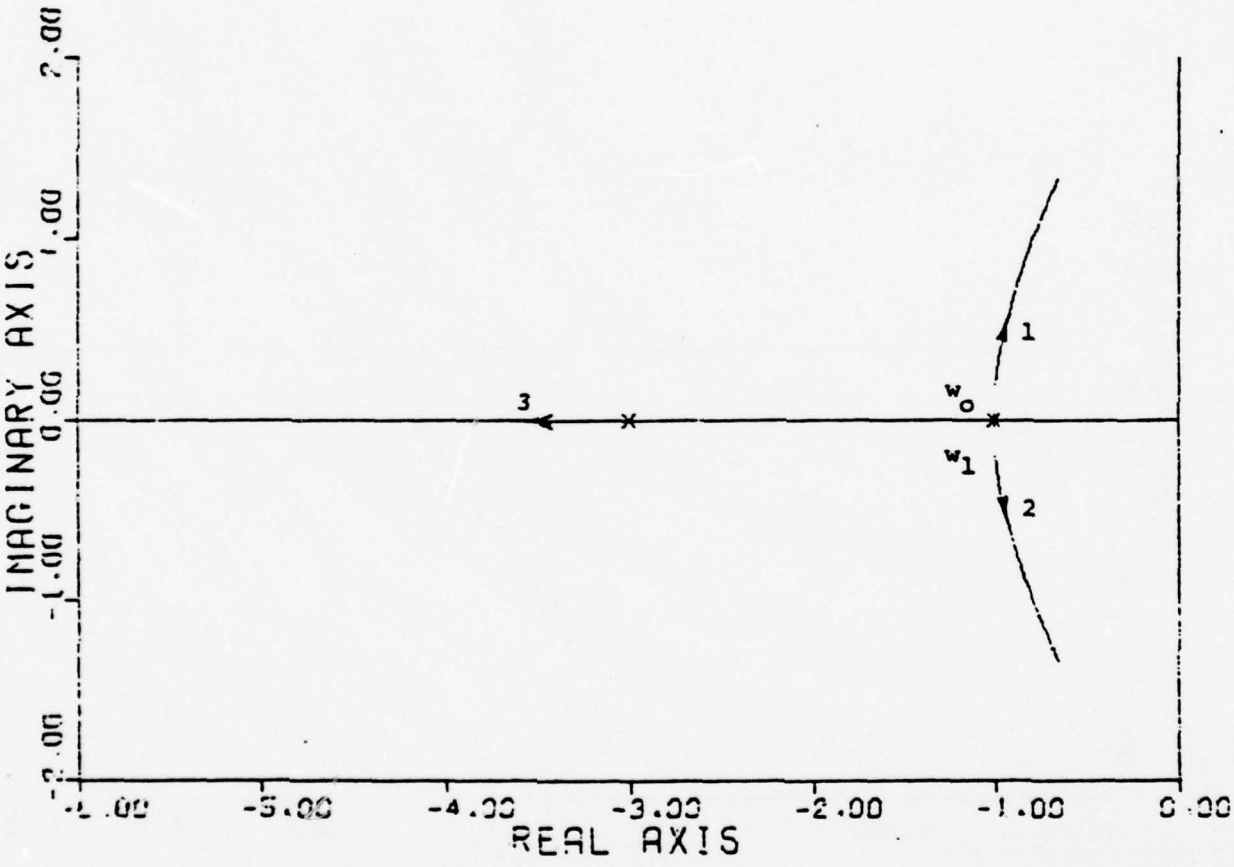


Fig. 5. $G(s)H(s) = \dfrac{K}{(s+3)(s+1)^2}$.

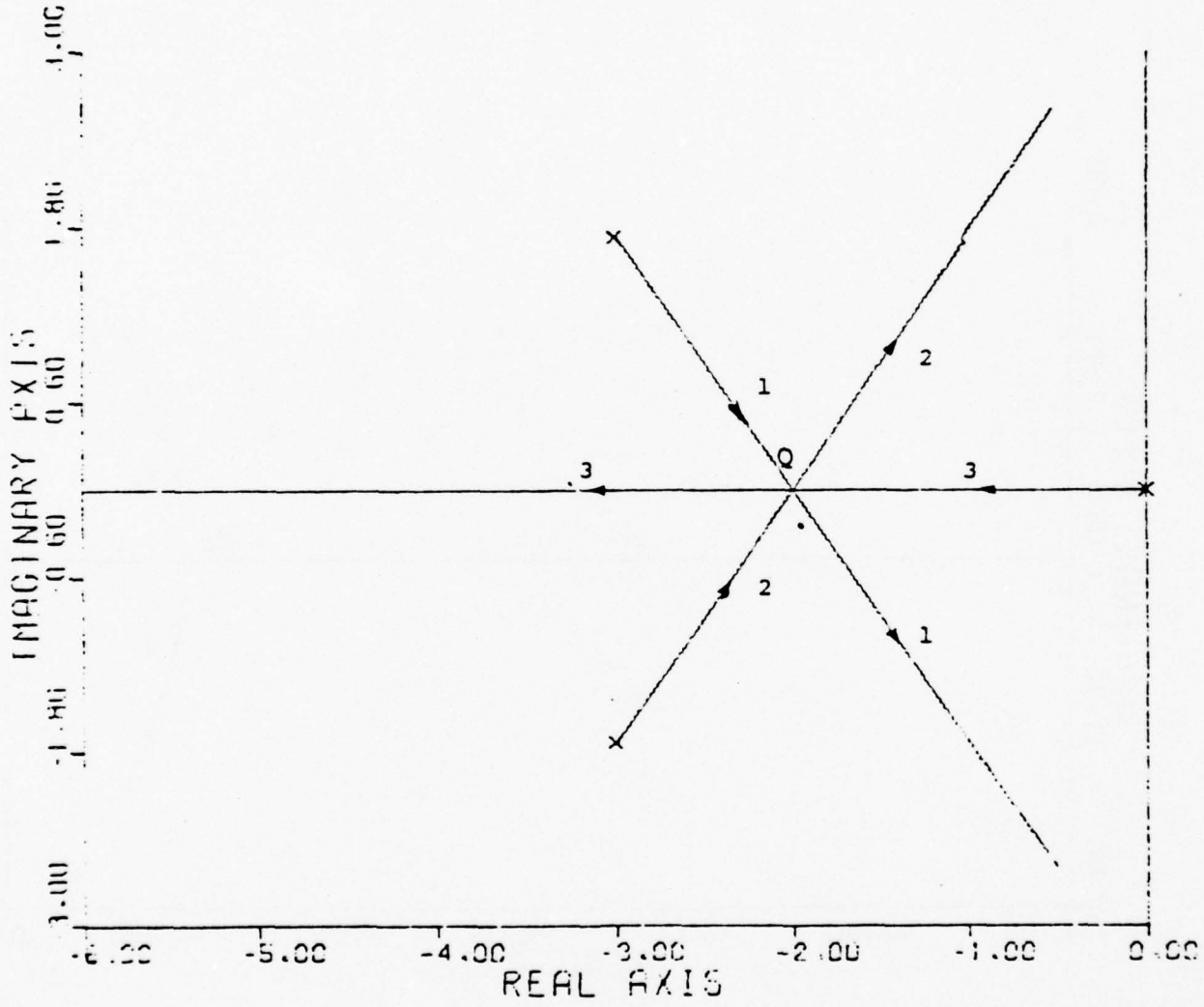


Fig. 6. $G(s)H(s) = \dfrac{K}{s(s + 3 + j\sqrt{3})(s + 3 - j\sqrt{3})}$.

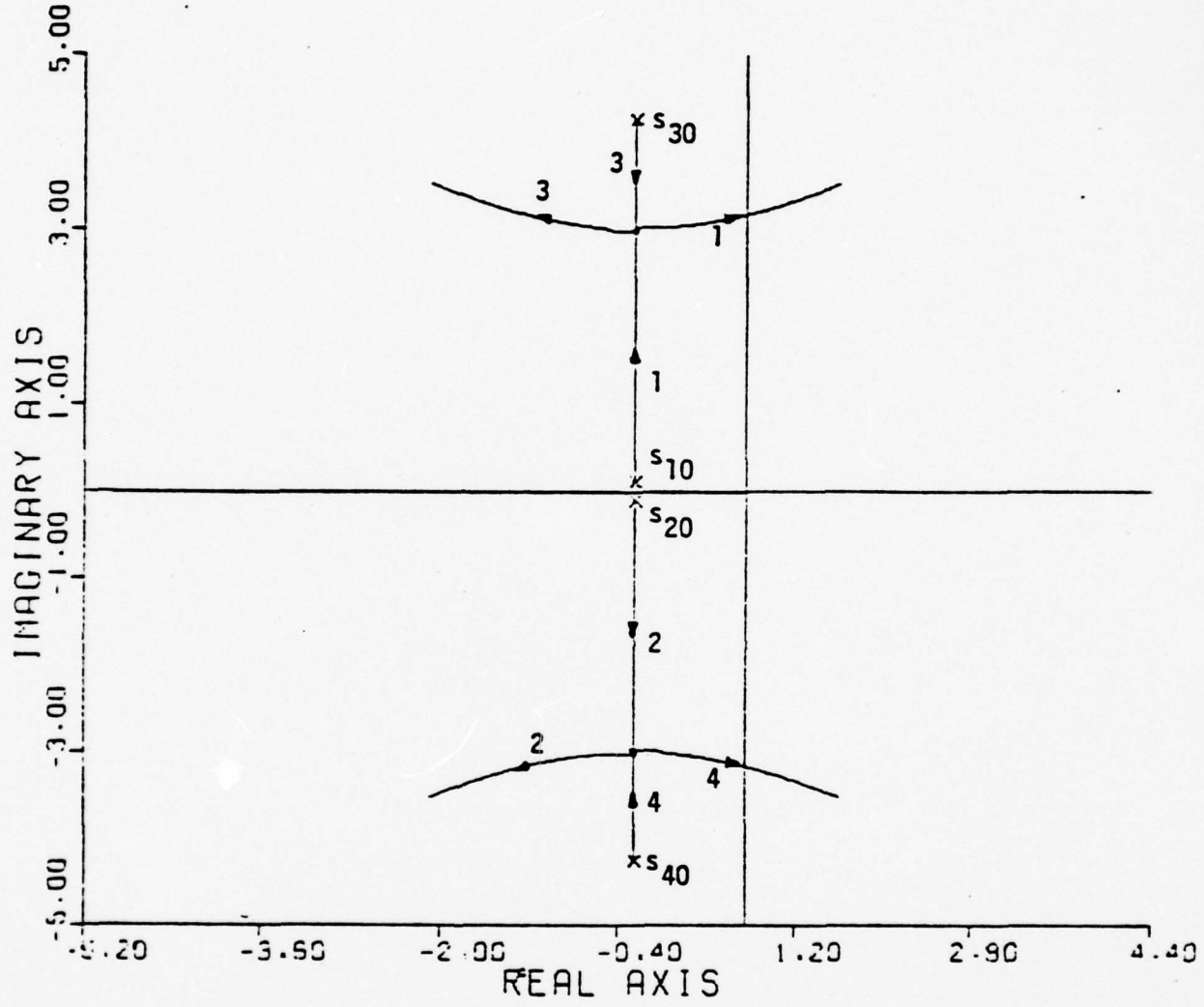


Fig. 7. $G(s)H(s) = \dfrac{K}{(s+1)^2(s+1+j\sqrt{18})(s+1-j\sqrt{18})}$.

RESEARCH

on

STOCHASTIC CONTROL AND ESTIMATION

R.B. Asher* and S.I. Marcus#

DEPARTMENT OF ELECTRICAL ENGINEERING

TEXAS TECH UNIVERSITY

*Prof. Asher is an associate investigator on the Stochastic Control and Estimation work unit for which the principal investigator is Prof. S.R. Liberty.

#Prof. Marcus is with the Department of Electrical Engineering, University of Texas at Austin, Austin, Tx.

Abstract

This paper considers the estimation of mixed rotational and doubly stochastic Poisson process observables. The estimation structure for the general nonlinear problem is obtained using Kushner's method for the development of the differential equations for the evolution of the conditional probability density functions.

Introduction

There are problems in which the measurement subsystems observing the state of a system are of fundamentally different types. The measurements may be a point process whereby the number and times of arrivals contain information about the state of the system. The measurements may contain additive noise or they may be rotational processes. Estimation of random point processes is considered in detail in the book by Snyder [1]. Estimation of rotational processes has been studied in detail in references [3-8], among others.

Many problems occur in actual applications in which there is a mix between the various types of measurement subsystems. For example, in reference [2], the problem of estimation of random thrusts in a solar electric propulsion spacecraft has a mixture of both doubly stochastic Poisson process measurements from a photon counting star tracker and additive noise measurements from Doppler signals. The theory for this mixture of estimation has been developed in reference [9] for the linear case and [11] for the nonlinear case, and independently by a different method in reference [2].

In this short paper, the problem of estimation of mixed rotational and point process observables is solved. In particular, the measurement subsystems consist of two different types. The first type is that of a doubly stochastic Poisson process and the second type is that of a rotational process. The equation for the evaluation of the conditional probability density function for the state

and the state modulo 2π is given, along with the functional forms for the density function. A general equation is given for the evolution of moments and from this a differential equation for the evolution of the conditional mean is given. Exponential Fourier series [8] is used to develop approximate estimations for the case when the intensity rate is a periodic process. These results are contained in the long version of this paper.

Problem Statement

This section contains the problem statement for optimal and suboptimal estimation of mixed rotational and point process observables. The system and measurement models are given.

Consider the probability space (Ω, A, P). The system equation is given as

$$dx(t) = f[x(t), t]dt + g[x(t), t]d\beta(t) \tag{1}$$

where $x: \Omega \to R$ and is Borel measureable, β is a Wiener process with unity variance parameter, $f: R' \times R' \to R'$ and is Borel measurable. One of the measurements subsystems observes the state x through the transformation which folds z(t) around the unit circle. This measurement process as given in [3] is either in the form

$$Z(t) = J[z(t)] \tag{3}$$

or

$$dZ(t) = \begin{bmatrix} \frac{-r(t)}{2}\,dt & dz(t) \\ -dz(t) & \frac{-r(t)}{2}\,dt \end{bmatrix} Z(t) \tag{4}$$

where

$$J[z(t)] = \begin{bmatrix} \cos z(t) & \sin z(t) \\ -\sin z(t) & \cos z(t) \end{bmatrix} \tag{5}$$

which maps the process onto the unit circle. This type of measurement process is explored in detail in the series of papers by Lo and Willsky [3,4,5].

The process x(t) is observed by another measurement subsystem which yields a counting measurement from a doubly stochastic Poisson process N(t) with a periodic intensity rate of the form

$$\lambda[x(t), t] \text{ where } \quad \lambda = R^1 \times R^1 \to R^1$$

The probability of n counts in an internal of time, conditioned on the intensity rate, is given as

$$\Pr\{ N^{(t)} - N^{(s)} = n \mid \lambda [x(\sigma), \sigma], s < \sigma \leq t\} =$$

$$\frac{(\int_s^t \lambda[x(\sigma), \sigma]\, d\sigma)^n \exp(-\int_s^t \lambda[x(\sigma, \sigma)] d\sigma)}{n!} \tag{6}$$

where N(t) is the total number of counts from t_o to t. It is assumed that the processes n_t and w_t are indeptendent.

Let T_t be the sub- -algebra of A generated by $Z^t \triangleq \{Z(\tau)^t, t_o \leq \tau \leq t\}$ and let $N_t \triangleq \{N(\tau), t_o \leq \gamma \leq t\}$. Since the information contained in the process Z is the same as the process $z^t \triangleq z(\tau), t_o \leq \tau \leq t$ (See reference [3]) due to the bijectivity of the Lie group mapping, the sub-σ-algebra generated by z^t is the same as that by Z^t. Let $B_t = T_t \vee N_t$ be the smallest σ-albegra containing T_t and N_t. Then the estimation problem to be solved is to obtain the estimates of x(t) given the σ-algebra B_t.

The next section contains the equations for the evolution of the conditional density functions, an equation for the evolution of moments, and the solution for the differential equation for the conditional mean.

Solution of the Optimal Estimation Problem

The conditional probability density function for x(t) given B_t may be written as

$$\underline{P}_x[(t)|\mathcal{B}_t] = \frac{E\{\exp\psi^t|x(t) = x\}P_x[x(t)]}{E\{\exp\psi^t\}} \tag{7}$$

where

$$\psi^t = -\tfrac{1}{2}\int_{t_o}^{t} \frac{h^2[x(\sigma), \sigma]}{r(\sigma)}\, d\sigma +$$

$$+ \int_{t_o}^{t} \frac{h[x(\sigma), \sigma]}{r(\sigma)} [Z^t(\sigma)dZ(\sigma)]_{12} \tag{8}$$

$$- \int_{t_o}^{t} \lambda[x(\sigma), \sigma]d\sigma + \int_{t_o}^{t} \text{In}\ \lambda[x(\sigma), \sigma]dN_\sigma$$

where $\int$ denotes an It_o integral and the last integral is a counting integral [1]. This may be proved by using the fact that, given x(t), the two measurment sybsystems are independent; the representation theorems of [1] and [3] can then be applied. The foled density may be written as

$$P_\theta[\theta(t)|\mathcal{B}_t] =$$

$$\sum_{k=-\infty}^{\infty} \frac{E\{\exp\psi^t|x(t) = \theta(t) + 2k\pi\}P_x(\theta(t) + 2k\pi)}{E\{\exp\psi^t\}} \tag{9}$$

In a similar manner as [2] or [3], the stochastic partial differential equations for the temporal evolution of the conditional density functions may be derived for $P[x(t)|\mathcal{B}_t]$ and $P[\theta(t)|\mathcal{B}_t]$.

Theorem 1: The stochastic parial differential equation for the evolution of $P_x[x(t)|\mathcal{B}_t]$ is

$$dP_x[x(t)|\mathcal{B}_t] = LP_x[x(t)|\mathcal{B}_t]dt + \frac{1}{r(t)}\{h[x(t), t] -$$

$$\hat{h}[x(t), t]dt\}\{[Z(t)^T dZ(t)]_{12} - \hat{h}[x(t), t]dt\}$$

$$.P_x[x(t)|\mathcal{B}_t] + \{\lambda[x(t), t] - \hat{\lambda}[x(t), t]\}$$

$$.\hat{\lambda}[x(t), t]^{-1}\{dN(t) - \hat{\lambda}[x(t), t]dt\}P_x[x(t)|\mathcal{B}_t] \tag{10}$$

where

$$L(\cdot) = \frac{\partial(.f)}{\partial x} + \tfrac{1}{2}\,\frac{\partial^2(.g)}{\partial x^2} \tag{11}$$

and

$$\hat{h}[x(t), t] = E\{h[x(t), t] | \mathcal{B}_t\}. \tag{12}$$

Proof: The proof may be obtained by using Kushner's method [10], and it follows as in reference [2].

The differential equation for $P[\theta(t)|\mathcal{B}_t]$ follows as in [3] and is

$$dP_\theta[\theta(t)|\ _t] = \sum_{n=-\infty}^{\infty} dP_x[\theta(t) + 2\pi n|\mathcal{B}_t]. \tag{13}$$

Theorem 2: Let $\phi[x(t)]$ be a twice continuously differentiable scalon function of the state $x(t)$ which is the solution of equation (1). Then the differential equation for the $\mathcal{B}_t$ conditional expectation of ϕ, $E^{\mathcal{B}_t}\{\phi[x(t)]\}$, is given as

$$\begin{aligned}
dE^{\mathcal{B}_t}\{\phi[x(t)]\} &= E^{\mathcal{B}_t}\{\phi_x[x(t)]\ f[x(t), t]\}dt + \\
&+ \tfrac{1}{2}\,E^{\mathcal{B}_t}\{g^2[x(t), t]\phi_{xx}[x(t)]\}dt + \\
&+ \{E^{\mathcal{B}_t}\{\phi[x(t)]h[x(t), t]\} - \\
&- E^{\mathcal{B}_t}\{\phi[x(t)]\ \}E^{\mathcal{B}_t}\{h[x(t), t]\}\} \\
&\cdot \frac{1}{r(k)}\,\{[Z(t)^T dZ(t)\ ]_{12} - \\
&- E^{\mathcal{B}_t}\{h[x(t), t]\}dt\} + \{E^{\mathcal{B}_t}\{\phi[x(t)] \\
&\cdot \lambda[x(t),t]\} - E^{\mathcal{B}_t}\{\phi[x(t)]\}\ E^{\mathcal{B}_t} \\
&\cdot \{\phi[x(t)]\}\ E^{\mathcal{B}_t}\{\lambda[x(t), t]\}\} \\
&\cdot E^{\mathcal{B}_t}\{\lambda[x(t), t]\}^{-1}\{dN(t) - \\
&- E^{\mathcal{B}_t}\{\lambda[x(t), t]\}dt\}
\end{aligned}$$

Proof: This follows from reference [2].

Theorem 3: Given the system and measurement subsystems as previously defined, the equation for $\hat{x}(t)$, $E\{x(t)|B_t\}$, may be written as

$$d\hat{x}(t) = E^{B_t}\{f[x(t), t]\}dt + \frac{1}{r(t)} E^{B_t}\{[x(t) -$$

$$- \hat{x}(t)]h[x(t), t]\} \{[Z(t)^T dZ(t)^T]_{12} -$$

$$- E^{B_t}\{h[x(t), t]\}dt\} + E^{B_t}\{[x(t) - \hat{x}(t)]$$

$$.\lambda[x(t), t]\} E^{B_t}\{\lambda[x(t), t]\}^{-1}\{dN(t) -$$

$$- E^{B_t}\{\lambda[x(t), t]\} dt\} .$$

The covariance of the estimation error may be written as

$$dP(t) = 2E^{B_t}\{(x(t) - \hat{x}(t)f[x(t), t]\}dt +$$

$$+ E^{B_t}\{g[x(t), t]^2\}dt + E^{B_t}\{(x(t) -$$

$$- \hat{x}(t))^2 (h[x(t), t] - E^{B_t}\{h[x(t), t]\})\}$$

$$.\{[Z(t)^T dZ(t)]_{12} - E^{B_t}\{h[x(t), t]\} dt\}$$

$$- \frac{1}{r(t)} E^{B_t}\{(x(t) - \hat{x}(t))h[x(t), t]\}^2 dt +$$

$$+ E^{B_t}\{(x(t) - \hat{x}(t))^2 (\lambda[x(t), t] -$$

$$E^{B_t}\{\lambda[x(t), t]\}\} E^{B_t}\{\lambda[x(t), t]\}^{-1}$$

$$.\{dN(t) - E^{B_t}\{\lambda[x(t), t]\}dt\} - E^{B_t}\{(x(t) -$$

$$- \hat{x}(t))\lambda[x(t), t]\}^2 E^{B_t}\{\lambda[x(t), t]\}^{-2} dN(t).$$

Proof: Theorem 2 may be used directly.

Since the estimator of Theorem 3 is infinite dimensional, it is necessary to consider approximate suboptimal estimators.

An approximate estimator structure may be obtained by using moment approximations, as in [1] and [12]. It will be assumed that the estimate is unbiased and the fourth order moments of the estimation error can be factored into products of second order moments similar to Gaussian moments. Furthermore, all remaining moments of the estimation error other than the second will be eliminated. With these assumptions the approximate estimation structure may be written as

$$d\hat{x}(t) = f[\hat{x}(t), t]dt + \frac{P(t)}{r(t)} \frac{\partial h}{\partial x}\Big|_{\hat{x}} \{[Z(t)^T dZ(t)]_{12}$$

$$- h[\hat{x}(t), t]dt\} + P(t) \frac{\partial \lambda}{\partial x}\Big|_{\hat{x}}$$

$$\cdot \{\lambda[\hat{x}(t), t]\}^{-1}\{dN(t) - \lambda[\hat{x}(t), t]dt\}$$

where the approximate estimation error covariance

$$dP(t) = \{2P(t) \frac{\partial f}{\partial x}\Big|_{\hat{x}} + g^2$$

$$- P(t)^2 \frac{\partial^2 [\hat{x}(t), t]}{\partial \hat{x}^2} - \frac{P(t)^2}{r(t)} (\frac{\partial h}{\partial x}|_{\hat{x}})^2\}dt +$$

$$+ P(t)^2 \frac{\partial^2 In\lambda}{\partial x^2}\Big|_{\hat{x}} dN(t)$$

Conclusion

This paper considers the problem of estimation using mixed observables when the observables are of two fundamentally different types. The first type is that of a rotational process and the second type is that of a doubly stochastic Poisson process. The estimation structure is obtained using the

differential equation for the conditional probability density function. A suboptimal estimator based on moment approximation is given for this problem.

Exponential Fourier series [8] will be used to develop additional approximate estimators for the case when the intensity rate is a periodic process. These results are contained in the long version of the paper.

References

1. D.L. Snyder, Random Point Processes, Wiley-Interscience Publishers, New York, 1975.

2. R.B. Asher, T.J. Eller, S.R. Robinson, M.D. Shackelford, and B.D. Tapley, "Mixed Observable Estimation of Random Thrust Errors for Solar Electric Propulsion Spacecraft," Proc. of the 1977 AIAA Guidance and Control Conf., Hollywood, Fl., Aug. 1977.

3. J.T-H. Lo and A.S. Willsky, "Estimation for Rotational Processes with One Degree of freedom-Part I: Introduction and Continuous Time Processes," IEEE Trans. on Auto. Contr., Vol. AC-20, No. 1, pp. 10-21, Feb. 1975.

4. A.S. Willsky and J. T-H. Lo, "Estimation for Rotational Processes with One Degree of freedom-Part II: Discrete-Time Processes," IEEE Trans. on Auto. Contr., Vol. AC-20, No. 1, pp. 22-30, Feb. 1975.

5. A.S. Willsky and J. T-H. Lo, "Estimation for Rotational Processes with One Degree of Freedom-Part III: Implementation," IEEE Trans. on Auto. Contr., Vol. AC-20, No. 1, pp. 31-33, Feb. 1975.

6. A.S. Willsky, "Fourier Series and Estimation on the Circle with Application to Synchronous Communication-Part I: Analysis," IEEE Trans. on Information Theory, Vol. IT-20, No. 5, pp. 577-583, Sept. 1974.

7. A.S. Willsky, "Fourier Series and Estimation on the Circle with Application to Synchronous Communications-Part II: IEEE Trans. on Inform. Theory, Vol. IT-20, No. 5, pp. 584- , Sept. 1974.

8. J. T-H. Lo, "Exponential Fourier Densities and Optimal Estimations and Detection on the Circle," IEEE Trans. on Inform. Theory, Vol. IT-23, No. 1, pp. 110-116, Jan. 1977.

9. I.B. Rhodes and D.L. Snyder, "Estimation and Control Performance for Space Time Point Process Observations," IEEE Trans. on Auto. Center, Vol. AC-22, No. 3, pp. 338-346, June 1977.

10. H.J. Kushner, "On the Dynamical Equations of Conditional Probability Density Functions with Applications to Optimal Stochastic Control", J. of Math. Anal. and Appl., Vol. 8, pp. 332-344, 1964.

11. M.V. Vaca and D.L. Snyder, "Estimation and Decision for Observations Derived from Martingales: Part 2, Applications," Biomedical Computer Laboratory Monograph No. 298, Washington Univ. Schhol of Medicine, St. Louis, Missouri, Sept. 1976.

12. A.H. Jazwinski, Stochastic Processes and Filtering Theory, Academic Press, New York, 1970.

RESEARCH

on

DECENTRALIZED CONTROL

R. Saeks*

DEPARTMENT OF ELECTRICAL ENGINEERING

TEXAS TECH UNIVERSITY

*Prof. Saeks is an associate investigator on the Decentralized control work unit for which the principal investigator is Prof. D. Gustafson.

Abstract

It is shown that the fixed modes of an interconnected dynamical system under decentralized control are precisely the uncontrollable and unobservable states of the individual system components. As such, the system can be stabilized by decentralized controllers if and only if its individual system components can be stabilized. Moreover, these conditions are shown to be equivalent to the conditions for stabilizing the system using a global controller.

Introduction

Given a linear system with partitioned inputs and outputs

$$\dot{X} = FX + \sum_{i=1}^{n} B^i u_i$$

$$y_i = C^i X \qquad i = 1,2, \ldots, n \tag{1.}$$

it is desired to design a family of dynamic decentralized controllers

$$\begin{aligned} Z_i &= S_i Z_i + R_i y_i \\ u_i &= Q_i Z_i + K_i y_i \end{aligned} \qquad i = 1,2, \ldots, n \tag{2.}$$

which place the poles (eigen values) of the resultant feedback system in prescribed locations. In its most general form the solution to this problem was given by Wang and Davison.[1,2] Their solution is formulated in terms of the (diagonally) fixed modes of the systems

$$\theta_d(F,B,C) = \Omega\, \lambda(F+BK_dC) \tag{3.}$$

Here, B and C are the matrices $B = \text{col}(B^i)$ and $C = \text{row}(C^i)$, respectively, $\lambda(M)$ denotes the set of eigen values of the matrix M, and the intersection is taken over the set of block diagonal (complex[1]) matrices K_d whose partition is comformable with the partitions of B and C. Using this concept of fixed modes, Wang and Davison[1,2] showed that the eigen values of the system can be placed in a prespecified open region of the complex plane using the dynamic decentralized controllers of equation 2, if and only if θ_d lies in that region. More precisely, they showed that θ_d represents the set of eigen values of F which cannot be moved by any family of decentralized dynamic controllers while all remaining eigen values of F can be arbitrarily placed by an appropriate choice of decentralized dynamic controllers.

If one observes that $F+BK_dC$ is just the state matrix for the given system with the static decentralized feedback matrix K_d the above result can be interpreted as a characterization of the eigen value placement properties of the system under dynamic decentralized control in terms of its eigen value placement properties under static decentralized control. Indeed, the theorem of Wang and Davison states that those eigen values which can be moved at all by static controllers can be arbitrarily placed by dynamic controllers whereas those eigen values which are fixed under all static controllers are also fixed by dynamic controllers.

[1]Precisely the same theory can be formulated for systems characterized by real matrices though in that case the arguments are complicated by the fact that one must work with pairs of complex conjugate eigen values to preserve reality.

Since the partitioning in equation 1 is arbitrary, the above described theorem can be applied to the classical case wherein the given system has only a single input and output in which case the fixed modes of the system are given by

$$4. \qquad \theta(F,B,C) = \bigcap \lambda(F + BKC)$$

where the intersection is now taken over arbitrary matrices, K, which are conformable with B and C. Of course, in this special case $\theta(F,B,C)$ reduces to the usual set of eigen values which are either uncontrollable or unobservable.[4] Moreover,

$$5. \qquad \theta(F,B,C) \subset \theta_d(F,B,C)$$

since the intersection used to define θ is taken over a larger set of matrices than that used to define θ_d. Equation 5 formalizes the intuitively obvious fact that a system which is "controllable" by a family of decentralized controllers is also "controllable" by a global (centralized) controller.

The purpose of the present paper is to show that equation 5 holds with equality in the case where F, B, and C represent the dynamics of an interconnected dynamical system[4] in which y_i and u_i denote the local inputs and outputs associated with a given system component. As such, in that special case the eigen values of the system can be placed in prespecified locations by decentralized dynamic controllers whenever they can be placed in the same locations by a global dynamic controller. Although the class of interconnected dynamical systems is considerably smaller than the class of decentralized systems studied by Wang and Davison, the design of the local controllers for the components of an interconnected dynamical system is the "physical problem" which usually motivates the

study of the general decentralized control problem. As such, we believe that the above result is significant.

The class of interconnected dynamical systems which we consider is characterized schematically in Figure 1 and mathimatically by the set of equations

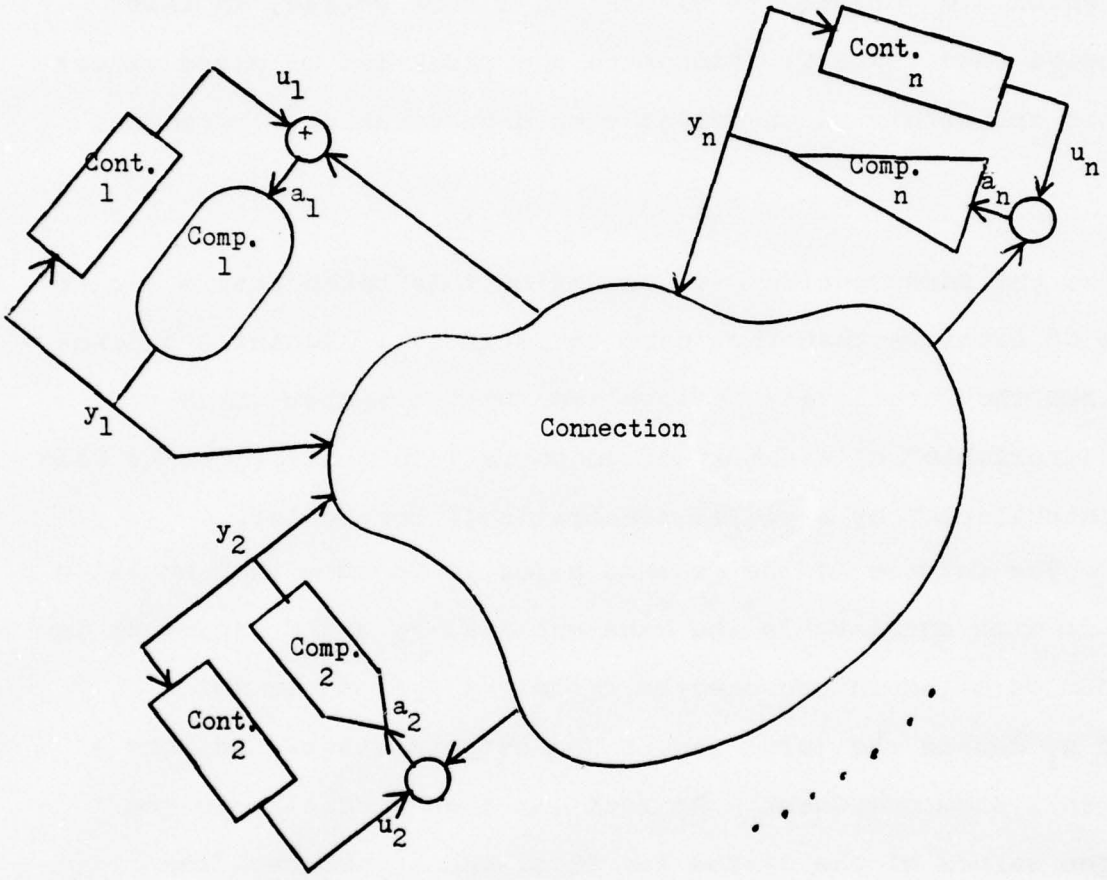


Figure 1. Interconnected dynamical system with local controllers.

$$
\begin{aligned}
\dot{X}_i &= A_i X_i + B_i a_i \\
y_i &= C_i X_i \qquad\qquad i = 1,2, \ldots, n \\
a_i &= \sum_{j=1}^{n} L^{ij} y_j + u_i
\end{aligned}
\tag{6.}
$$

Here the first two equations represent the dynamics of the ith component, whereas, the third equation defines the interconnection structure in which the input to the ith component is taken to be a linear combination of the outputs of the various components (including the ith) and an external control. The fact that the control inputs, u_i, are not multiplied by compansator matrices, implies that the local controllers, given by equation 2, have full access to the inputs of the individual system components, and, similarly, the output of the individual system components is fully accessible to the controllers.

For notational brevity equation 6 may be restated in block matrix form as

$$
\begin{aligned}
\dot{X} &= AX + Ba \\
y &= CX \\
a &= Ly + u
\end{aligned}
\tag{7.}
$$

where $X = \text{col}(X_i)$, $a = \text{col}(a_i)$, $y = \text{col}(y_i)$ $u = \text{col}(u_i)$, $A = \text{diag}(A_i)$, $B = \text{diag}(B_i)$, $C = \text{diag}(C_i)$ and $L = \text{mat}(L^{ij})$. Combining these into a single equation for the overall composite system we obtain the composite state model

$$
\begin{aligned}
\dot{X} &= FX + Bu \\
y &= CX
\end{aligned}
\tag{8.}
$$

where

$$F = A + BLC \tag{9.}$$

Moreover, upon observing that

10. $$Bu = \sum_{i=1}^{n} B^i u_i$$

and

11. $$y_i = C^i X \qquad i = 1, 2, \ldots, n$$

where $B^i = \text{col}(0,0 \ldots, 0, B_i, 0, \ldots, 0)$ and $C^i = \text{row}(0, 0, \ldots, 0, C_i, 0, \ldots, 0)$ we see that equation 8 naturally decomposes into the form of the decentralized control problem of equation 1. In equations 8, however, the B^i and C_i matrices take on a special form, whereas, they are arbitrary in equation 1. Indeed, it is this special form which yields the desired equality in equation 5. Intuitively, this implies that the ith local controller may drive only the state of the ith system component though that state may, in turn, drive the remainder of the system through the connection equations. Similarly, the ith local controller may observe only the state of the ith component with the remaining components being observed only indirectly through the state of the ith component.

Main Theorem:

Using the above notation, our main theorem may be stated as follows.

THEOREM: For the system of equation 8

12. $$\theta_d(F,B,C) = \theta(F,B,C) = \theta(A,B,C) = \bigcup_i \theta(A_i,B_i,C_i)$$

Proof: To show that $\emptyset(F,B,C) = \emptyset(A,B,C)$ we simple observe that

13. $$\lambda(F+BKC) = \lambda(A+B(L_{11}+K)C) = \lambda(A+BK'C)$$

where $K' = L_{11} + K$. As such, the same set of matrices are spanned if one takes the intersection of the $\lambda(F+BKC)$ over K or the intersection of the $\lambda(A+BK'C)$ over K' and hence $\theta(F,B,C) = \theta(A,B,C)$. Moreover, since A, B, and C are block diagonal $\theta(A,B,C)$ is just the union of the fixed modes associated with each block. Given 5 to prove the validity of the first equality of 2, it suffices to show that $\theta_d(F,B,C) \subset \theta(F,B,C)$. For this purpose, we desire to show that if λ is not in $\theta(F,B,C)$ then it is not in $\theta_d(F,B,C)$. Initially, we assume that A,B, and C are partitioned as 2 by 2 matrices, the general case following therefor by induction. If λ is not in $\theta(F,B,C)$ then there exists a K (dependent on λ) such that $\det(\lambda 1 - (F+BKC)) \neq 0$ and we desire to construct a block diagonal K_d, also dependent on λ, such that $\det(\lambda 1 - (F+BK_dC)) \neq 0$. To this end we write out the matrix F+BKC in partitioned form and expand its determinant via the formula for the determinant of a 2 by 2 partitioned matrix[3] obtaining

$$0 \neq \det(\lambda 1 - (F+BKC)) = \det(\lambda 1 - (A + B(L+K)C)$$

14.

$$= \det \left[\begin{array}{c|c} \lambda 1-A_1-B_1L^{11}C_1-B_1K^{11}C_1 & -B_1L^{12}C_2-B_1K^{12}C_2 \\ \hline -B_2L^{21}C_1-B_2K^{21}C_1 & \lambda 1-A_2-B_2L^{22}C_2-B_2K^{22}C_2 \end{array}\right]$$

$$= \det\,(\lambda 1-A_1-B_1L^{11}C_1-B_1K^{11}C_1)\det[\lambda 1-A_2-B_2L^{22}C_2-B_2K^{22}C_2$$

$$(E_2L^{21}C_1+B_2K^{21}C_1)(\lambda 1-A_1-B_1L^{11}C_1-B_1K^{11}C_1)^{-1}(B_1L^{12}C_2+B_1K^{12}C_2)]$$

$$= \det\,(\lambda 1-A_1-B_1L^{11}C_1-B_1K^{11}C_1)\det[\lambda 1-A_2-B_2L^{22}C_2-B_2\underline{K}^{22}C_2$$

$$-B_2L^{21}C_1(\lambda 1-A_1-B_1L^{11}C_1-B_1K^{11}C_1)^{-1}B_1L^{12}C_2]$$

where

$$\underline{K}^{22} = K^{22}+K^{21}C_1(\lambda 1-A_1-B_1L^{11}C_1-B_1K^{11}C_1)^{-1}B_1L^{12}C_2$$

15.
$$+ L^{21}C_1(\lambda 1-A_1-B_1L^{11}C_1-B_1K^{11}C_1)^{-1}B_1K^{12}C_2$$

$$+ K^{21}C_1(\lambda 1-A_1-B_1L^{11}C_1-B_1K^{11}C_1)^{-1}B_1K^{12}C_2$$

16.
$$L = \left[\begin{array}{c|c} L^{11} & L^{12} \\ \hline L^{21} & L^{22} \end{array}\right]$$

and

17.
$$K = \left[\begin{array}{c|c} K^{11} & K^{12} \\ \hline K^{21} & K^{22} \end{array}\right]$$

are partitioned to be conformable with A, B, and C. Now, if we define K_d via

18.
$$K_d = \left[\begin{array}{c|c} K^{11} & 0 \\ \hline 0 & \underline{K}^{22} \end{array}\right]$$

and compute $\det(\lambda 1-(F+BK_dC))$ we obtain

$$\det(\lambda 1-(F+BK_dC)) = \det(\lambda 1-(A + B(L+K_d)C)$$

19.
$$= \det\left[\begin{array}{c|c} \lambda 1-A_1-B_1L^{11}C_1-B_1K^{11}C_1 & -B_1L^{12}C_2 \\ \hline -B_2L^{21}C_1 & \lambda 1-A_2-B_2L^{22}C_2-B_2\underline{K}^{22}C_2 \end{array}\right]$$

$$= \det(\, 1-A_1-B_1L^{11}C_1-B_1K^{11}C_1)\det[\lambda 1-A_2-B_2L^{22}C_2-B_2\underline{K}^{22}C_2$$

$$-B_2L^{21}C_1(\lambda 1-A_1-B_1L^{11}C_1-B_1K^{11}C_1)^{-1}B_1L^{12}C_2]$$

$$= \det(\lambda 1 - (F+BKC)) \neq 0$$

Thus there is at least one block diagonal K_d matrix such that λ is not an eigen value of $F+BK_dC$ showing that λ is not in $\theta_d(F,B,C)$.

To extend the above argument from 2 by 2 partitioned matrices to n by n partitioned matrices we repete the above construction n-1 times as follows. Given an n by n

20. $$K = \left[\begin{array}{c|cc|c} K^{11} & K^{12} & K^{13} & \cdots \; K^{1n} \\ \hline K^{21} & K^{22} & K^{23} & K^{2n} \\ K^{31} & K^{32} & K^{33} & K^{3n} \\ \vdots & & & \vdots \\ \hline K^{n1} & K^{n2} & K^{n3} & \cdots \; K^{nn} \end{array}\right]$$

such that $\det(\lambda 1-F+BKC))\neq 0$ it is partitioned into a 2 by 2 matrix as shown by the double line whence the above argument is employed to formulate a matrix

21. $$\underline{K} = \left[\begin{array}{c|cc|c} K^{11} & 0 & 0 & \cdots \; 0 \\ \hline 0 & K^{22} & K^{23} & K^{2n} \\ 0 & K^{32} & K^{33} & K^{3n} \\ \vdots & & & \vdots \\ \hline 0 & K^{n2} & K^{n3} & \cdots \; K^{nn} \end{array}\right]$$

such that $\det(\lambda 1-(F+B\underline{K}C))\neq 0$. This matrix is then repartitionen

into a new 2 by 2 matrix as shown by the double line in equation 21 and the process is repeated. Since the 1-1 entry in the partitioned matrix is not affected by the process this results in a new matrix of the form

22. $$\underline{\underline{K}} = \left[\begin{array}{c|c|c|c|c} \underline{\underline{K}}^{11} & 0 & 0 & \cdots & 0 \\ \hline 0 & \underline{\underline{K}}^{22} & 0 & & 0 \\ \hline 0 & 0 & \underline{\underline{K}}^{33} & & \underline{\underline{K}}^{3n} \\ \hline \vdots & & & & \vdots \\ \hline 0 & 0 & \underline{\underline{K}}^{n3} & \cdots & \underline{\underline{K}}^{nn} \end{array}\right]$$

such that $(\lambda 1-(F+B\underline{\underline{K}}C) \neq 0$. Repeating the process n-1 times eventually results in a block diagonal matrix K_d such that $\det(\lambda 1-(F+BK_dC) \neq 0$ showing that if λ is not in $\theta(F,B,C)$ then it is also not in $\theta_d(F,B,C)$ thereby verifying that $\theta_d(F,B,C)$ is contained in $\theta(F,B,C)$ and completing the proof of the theorem.

Given the theorem, for the class of interconnected dynamical systems, local control is just as good a global control from the point of view of pole placement. From the point of view of optimal control, however, global control will, in general, still be superior since it gives one a greater range of options.[4]

References

1. Wang, S.H., and E.J. Davison, "On the Controllability and Observability of Composite Systems", IEEE Trans. on Auto. Cont., Vol. AC-18, pp. 74-75, (1973).

2. Davison, E.J., and S.H. Wang, "New Results on the Controllability and Observability of General Composite Systems", IEEE Trans. on Auto. Cont., Vol. AC-20, pp. 123-128, (1975).

3. Gantmacher, F. R., The Theory of Matrices, (2 vols.), New York, Chelsea, 1959.

4. Saeks, R., and R.A. DeCarlo, Interconnected Dynamical Systems, New York, Marcel Dekker, (to appear).

RESEARCH

on

MATHEMATICAL SYSTEM THEORY

J. Murray*

DEPARTMENT OF ELECTRICAL ENGINEERING

TEXAS TECH UNIVERSITY

*Dr. Murray is a Research Associate working on the Mathematical System Theory work and under R. Saeks who is the principal investigator.

Introduction:

The subject of two-dimensional digital filters has received considerable attention of late: in particular, two-dimensional spectral factorization has been treated in a number of papers - it is considered in great detail in reference [1]. The major problem which arises is that in general the spectral factors of a rational transfer function are not rational: some further processing, such as truncation and smoothing, is usually employed to yield approximate rational factors. It is, therefore, somewhat surprising that the class of rational functions for which a rational spectral factorization exists does not seem to have been investigated. In this paper, we give two sets of conditions which must be satisfied by such functions (theorems 1 and 3); a converse is given which may be applied to the numerator and denominator polynomials separately. Now, the polynomial spectral factors (when they exist) of a given polynomial are minimum- and maximum-phase polynomials; conversely, every such polynomial gives rise to trivial spectral factors. Motivated by this, we apply the results of theorems 1 and 3 to the particular case of minimum-phase polynomials (i.e., polynomials without zeros in the unit polydisc).

In this context, the main consequences of the results of this paper may be broadly outlined as follows:

i) A given polynomial has exactly the same amplitude response as a minimum-phase polynomial if and only if the classical one-variable method (of factoring the original polynomial into a product of two polynomials devoid of zeros in certain regions) can be applied. (This result is in fact implicit in [1], but does not appear to have been explicitly stated in the literature). The corresponding statement for minimum-phase, stable rational functions is false, however.

ii) If the conditions given in theorems 1 and/or 3 are not satisfied, then not alone is there no minimum-phase stable rational function having exactly the same amplitude response as the original; the original amplitude response can not even be approximated arbitrarily well by minimum-phase stable rational functions. This follows from the fact that the conditions in theorems 1 and 3 are conditions on the amplitude response which are preserved under any reasonable kind of convergence.

iii) The conditions in theorem 3 are easily visualized and surprisingly stringent; they require essentially that the gain of the filter, averaged over certain directions in the frequency plane, have no variation in a perpendicular direction. (see the discussion following theorem 3). This gives extremely severe restrictions on the amplitude response of minimum-phase FIR filters, minimum-phase stable IIR filters, and the denominator polynomial of arbitrary stable IIR filters.

iv) It has been pointed out by Bose [9] and Woods [10], and again is implicit in [1], that there exist purely recursive filters whose amlitude responses are not realizable as the amplitude response of any stable purely recursive filter, and that consequently any stabilization method which attempts to match the amplitude response of the original filter is doomed to failure. The restrictions referred to in iii), above, reinforce this conclusion and identify the precise properties of the examples in [9] and [10] which make stabilization impossible.

Definitions and Notation:

Our notation will follow that in [2]; we repeat it here for convenience. For simplicity we restrict ourselves throughout to two dimensions, although

there does not appear to be any difficulty in extending the results to higher dimensions. Thus all functions are assumed throughout to be rational functions of two complex variables unless otherwise stated; we further exclude the zero function. Two-dimensional complex space will be denoted by $\mathbb{C}^2$, i.e, $\mathbb{C}^2 = \{(Z_1, Z_2) | Z_1 \text{ and } Z_2 \text{ are complex numbers}\}$. The open unit polydisc will be denoted by U^2, i.e.,

$$U^2 = \{(Z_1,Z_2) \in \mathbb{C}^2 \mid |Z_1| < 1 \text{ and } |Z_2| < 1\}$$

and its closure will be denoted by $\overline{U}^2$:

$$\overline{U}^2 = \{(Z_1, Z_2) \in \mathbb{C}^2 \mid |Z_1| \leq 1 \text{ and } |Z_2| \leq 1\}$$

The distinguished boundary of the unit polydisc will be denoted by T^2:

$$T^2 = \{(Z_1, Z_2) \in \mathbb{C}^2 \mid |Z_1| = 1 \text{ and } |Z_2| = 1\}$$

The frequency response of the filter whose transer function is $f(Z_1,Z_2)$ is simply the restriction of f to T^2. We will find it convenient to denote this restriction by f*.

The one-dimensional sets corresponding to the above are:

$$U = \{Z \in \mathbb{C} \mid |Z| < 1\}$$

$$\overline{U} = \{Z \in \mathbb{C} \mid |Z| \leq 1\}$$

$$T = \{Z \in \mathbb{C} \mid |Z| = 1\}$$

We need one further subset of $\mathbb{C}^2$:

$$V^2 = \{(Z_1,Z_2) \in \mathbb{C}^2 \mid |Z_1| > 1 \text{ and } |Z_2| > 1\}.$$

By the Fourier coefficients of a function $h(\theta_1,\theta_2)$ defined on T^2 we mean the numbers

$$a_{mn} = \frac{1}{4\pi^2} \int_0^{2\pi} \int_0^{2\pi} h(\theta_1,\theta_2)\, e^{-j(m\theta_1 + n\theta_2)} d\theta_1\, d\theta_2.$$

Finally, let us state precisely what we mean by the term spectral factorization. Several different forms of spectral factorization are treated in [1]; here we will be concerned only with the simplest form: if f is a rational function, it will be said to have a (rational, quarter-plane) spectral factorization if $f = f_1 f_2$ where f_1 and f_2 are rational functions, f_1 has no poles or zeros in U^2, and f_2 has no poles or zeros in V^2. Several comments are in order concerning this definition:

i) By "rational" we mean only "finite-order"; i.e., the functions are assumed to be expressible as the quotient of two (finite-order) polynomials.

ii) The quarter-plane property enters only in connection with the regions in which the factors are assumed to be zero- and pole-free; in particular, if f has no poles or indeterminacies on T^2, and has a quarter-plane spectral factorization, then there is a quarter-plane causal, stable filter whose amplitude response is equal to $|f^*|$.

iii) It would possibly be more natural to work with $\overline{U^2}$ and $\overline{V^2}$ rather than U^2 and V^2 (especially when considering stability). However, to do so would complicate the statements of the theorems considerably, and it is usually clear whether or not the results will hold with $\overline{U^2}$ and $\overline{V^2}$ in place of U^2 and V^2. (One needs only to check for zeros and poles on T^2). In general, if the "closed" version is not obvious, it is not true; $1 - Z_1 Z_2$ will serve as a counterexample in all such cases.

iv) To simplify the statements of the theorems, the definition has been given in terms of the rational function f itself, rather than the spectral function $|f^*|^2$; however, the conditions given in the theorems actually involve only $|f^*|^2$.

v) We note that V^2 is defined to be a subset of $\mathbb{C}^2$; thus the behaviour of functions at infinity is irrelevant to our purposes.

Spectral Factorization:

Our first criterion for the existence of rational spectral factors is very much in the spirit in which spectral factorization is treated in [1]; it is a trivial consequence of theorem 5.4.7 in [2].

Theorem 1:

If a rational function f on $\mathbb{C}^2$ has a rational spectral factorization then the Fourier coefficents a_{mn} of log $|f^*|$ are zero for all pairs of integers (m,n) such that $m \neq 0$, $n \neq 0$, and m and n have different signs - that is, for all integer points in the second and fourth quadrants The converse is true for polynomial f.

As mentioned above, this criterion involves only the absolute value of f; it follows that the existence of spectral factors imposes restrictions on the amplitude response of a two-dimensional filter - in contrast with the situation in one dimension. The above criterion, however, does not present these restrictions in an easily visualized form - for instance, it is difficult to guage exactly how severe the restrictions are. For this reason, we next present conditions which are stated in terms of the log-amplitude response itself, rather than its Fourier coefficients. This result takes an approach which seems to differ substantially from those previously known; it gives easily visualized necessary conditions on those rational functions which admit a rational spectral factorization. Before we state this theorem, however, we first present a simple result which will

be used in the proof, and is also of separate interest; one of its consequences is that when rational spectral factors exist, the usual one-dimensional stabilization method (for unstable denominator polynomials) can be used.

Theorem 2:

If the rational function f admits a rational spectral factorization, then there is a rational function $\tilde{f}$ (with deg $\tilde{f} \leq$ deg f) such that

$$|\tilde{f}^*| = |f^*|$$

and $\tilde{f}$ has no poles or zeros in U^2.

Again, the converse holds for polynomial f.

Thus, if the denominator polynomial of an unstable filter has polynomial spectral factors, there is a stable filter of at most the same order with the same amplitude response (provided the polynomial has no zeros on T^2).

Again most of the proof is contained in [2]; we fill in the details here: suppose f has rational spectral factors, then $f = f_1 \frac{P}{Q}$ where f_1 has no poles or zeros in U^2 and P and Q are polynomials without zeros in V^2.

Let $$\tilde{P} = Z_1^m Z_2^n \overline{P}(1/Z_1, 1/Z_2),\ Z_2 \neq 0,\ Z_1 \neq 0$$

where m is the degree of P in Z_1, n is the degree of P in Z_2, and $\overline{P}$ is the polynomial whose coefficients are the complex conjugates of the coefficients of P. Clearly $\tilde{P}$ is a polynomial of degree less than or equal to the degree of P, and so is also defined for $Z_1=0$ and $Z_2=0$. Now if $\tilde{P}(Z_1,Z_2) = 0$ for $Z_1 \neq 0$ and $Z_2 \neq 0$, then $\overline{P}(1/Z_1, 1/Z_2) = 0$; this implies that either

$$|1/Z_1| \leq 1 \text{ or } |1/Z_2| \leq 1 \text{ (since P has no zeros in } V^2)$$

and so either

$$|Z_1| \geq 1 \text{ or } |Z_2| \geq 1, \text{ i.e., } (Z_1, Z_2) \notin U^2$$

Thus the only possible zeros of $\tilde{P}$ in U^2 are for $Z_1 = 0$ or $Z_2 = 0$. But by standard results in the theory of several complex variables [8], if the zero-set were nonempty, this would imply that either Z_1 or Z_2 was a factor of $\tilde{P}$, which is impossible by our choice of m and n. Thus $\tilde{P}$ has no zeros in U^2. Finally, on T^2

$$|\tilde{P}(Z_1,Z_2)| = |Z_1^m \bar{Z}_2^{\bar{m}} \bar{P}(1/Z_1,1/Z_2)| = |\bar{P}(\bar{Z}_1,\bar{Z}_2)| = |P(Z_1,Z_2)|.$$

$\tilde{Q}$ is defined similarly and has similar properties. Then

$$\tilde{f} = f_1 \frac{\tilde{P}}{\tilde{Q}}$$

clearly has the required properties.

Conversely, suppose f is any polynomial for which there is a rational function $\tilde{f}$ without poles or zeros in U^2 such that

$$|\tilde{f}^*| = |f^*|$$

then $f/\tilde{f}$ is rational and analytic in U , and

$$|(f/\tilde{f})^*| = 1$$

Thus by theorems 5.2.5 and 5.2.6 in [2], $f/\tilde{f} = P/Q$ where P and Q are polynomials, P has no zeros in V^2, and Q has no zeros in U^2. Then

$$f = P\tilde{f}/Q$$

gives a rational (in fact, polynomial) spectral factorization of f.

The Second Criterion:

Our second set of conditions for the existence of a rational spectral factorization is given in the following:

Theorem 3:

If a rational function f on $\mathbb{C}^2$ admits a rational spectral factorization, then

$$\frac{1}{2\pi} \int_0^{2\pi} \log |f(e^{jm\theta}, e^{j(n\theta + \Psi)})| d\theta$$

is a constant independent of Ψ, $(0 \leq \Psi < 2\pi)$, for all integers $m > 0$ and $n > 0$.

Again, these conditions depend only on the amplitude response of f. The simplest condition is that for m = 1 and n = 1; it can be easily visualized by drawing two adjacent squares in the $\theta_1\ \theta_2$ - plane on which the amplitude response is defined (the frequency response extends to the entire $\theta_1\theta_2$ - plane by periodicity), and drawing lines L_i with slope 1 and length $2\pi\sqrt{2}$ on these squares; see figure 1.

Then the condition for m = 1, n = 1 can be restated as: the "average" amplitude of the function f along the line L_i is a constant - that is, it is as independent of the particular line L_i chosen. ("Average" here is to be understood as the geometric mean of the amplitude, or the arithmetic mean of the log-amplitude). Alternatively, we may say that the average level of the amplitude over any line of slope 1 and of length $2\pi\sqrt{2}$ is independent of the position of the line in the $\theta_1\theta_2$-plane. (For example, we could vary the L_i over the dotted square in the direction $\underline{\hat{n}}$). The conditions for higher m and n have a similar interpretation, with a slope of n/m instead of 1, and length $2\pi\sqrt{m^2 + n^2}$ instead of $2\pi\sqrt{2}$; clearly, if m and n are not relatively prime, the corresponding condition is superfluous.

This theorem then gives a striking limitation on the amplitude response of a rational function which admits a rational spectral factorization; even the simplest of the conditions (that for n=m=1) implies that

such a function can not accurately approximate an amplitude which has large variations in overall level in the direction $\underline{\hat{n}}$ shown in figure 1.

Proof of Theorem 3:

In view of theorem 2, it suffices to prove this under the assumption that f has no poles or zeros in U^2. This assumption implies that f has a holomorphic logarithm in U^2. Then, for any integers $m > 0$, $n > 0$ and any real number Ψ,

$$\log f(Z^m, Z^n e^{j\Psi})$$

is a holomorphic function of one complex variable for $Z \in U$. Thus

$$\mathrm{Re}(\log f(Z^m, Z^n e^{j\Psi}))$$

is a harmonic function in U, and so by the mean-value property of harmonic functions

$$\frac{1}{2\pi} \int_T \mathrm{Re}\ (\log f(Z^m, Z^n e^{j\Psi}))\ d\theta = \mathrm{Re}\ (\log f(o^m, o^n e^{j\Psi}))$$

$$\text{i.e., } \frac{1}{2\pi} \int_0^{2\pi} \mathrm{Re}\ (\log f(e^{jm\theta}, e^{j(n\theta+\Psi)}))\ d\theta = \mathrm{Re}(\log f(o,o))$$

But $\mathrm{Re} \log w = \log |w|$ for $w \neq o$, and so

$$\frac{1}{2\pi} \int_0^{2\pi} \log |f(e^{jm\theta}, e^{j(n\theta + \Psi)})|\ d\theta = \log |f(o,o)|$$

and the right-hand side is independent of Ψ (and, incidentally, of m and n also).

An obvious question which arises is the extent to which the converses of these results hold. In fact, the converse of Theorem 3 holds for polynomials, and modified converses of both Theorems 1 and 3 hold even for rational functions. The modification takes the following form: If the Fourier coefficients of $\log |f^*|$ (where f is a rational function) vanish

for $mn < o$, then there is a rational function $\tilde{f}$ with rational spectral factors, (equivalently, a rational function without poles or zeros in U^2), such that $|\tilde{f}^*| = |f^*|$. (A similar statement holds for Theorem 3). However, the proofs of these converses involve some technical analytic details, and so are relegated to an appendix.

The modification in the above converses lies, of course, in the fact that we cannot conclude that f itself has rational spectral factors; thus there are some rational functions which can be stabilized without changing the amplitude response but to which the classical 1-variable factorization technique cannot be applied. A simple example of this is the function

$$f(Z_1, Z_2) = \frac{Z_1 + Z_2 - 1}{Z_1 + Z_2 - Z_1 Z_2}$$

Here, $|f^*|$ is identically 1, and so has trivial spectral factors; but f itself clearly does not.

Although the converses of theorems 1 and 3 are proved in the appendix, there is another result related to the converse of Theorem 3; by strengthening the conditon for m=n=1 alone, we can get a stronger converse for polynomials. Before we state this converse, however, we first give a stability criterion (used in the proof of the converse) which, although previously known, [3], has not appeared in the engineering literature. Although not as sharp (in terms of dimension) as some other known criteria [4], it has two advantages which make it useful for theoretical purposes: first, it is given in terms of a one-parameter family of discs without the lower-dimensional test in [5]; and second, unlike most other stability tests, which conclude the nonvanishing of a polynomial on $\overline{U}^2$ from its nonvanishing on some subset of $\overline{U}^2$ which contains $\overline{T}^2$, this test allows the polynomial to vanish at some points in $\overline{T}^2$, but concludes only that the polynomial does

not vanish on U^2. The criterion is:

Theorem 4:

Suppose a polynomial f has no zeros in the set

$$\{(Z_1, Z_2) \in U^2 \mid |Z_1| = |Z_2|\};$$

then f has no zeros in U^2.

This is proved in a much more advanced context in [3]; however, it can also be easily proved by applying one of the criteria in [4] to the polydiscs

$$\overline{U}_r^2 = \{(Z_1,Z_2) \in \mathbb{C}^2 \mid |Z_1| \leq r, |Z_2| \leq r\}$$

for $0<r<1$.

For the hypotheses imply that f has no zeros on the distinguished boundary of $\overline{U}_r^2$ (for $0<r<1$), and none on the set

$$\{(Z_1,Z_2) \in \mathbb{C}^2 \mid Z_1=Z_2\} \cap \overline{U}_r^2$$

Thus by theorem 5 in [4], f has no zeros in $\overline{U}_r^2$ for any $r < 1$, and so f has no zeros in U^2.

We can now state and prove the partial converse to theorem 3.

Theorem 5:

If f is a polynomial with the property that

$$\frac{1}{2\pi} \int_0^{2\pi} \log \left| f(e^{j\theta}, e^{j(\theta+\Psi)}) \right| d\theta = \log |f(0,0)|, \text{ for } 0 \leq \Psi < 2\pi,$$

then f has no zeros in U^2.

Thus we strengthen the condition for m=1 and n=1 in theorem 3 by specifying that the constant in question is to be $\log |f(0,0)|$: it then follows not only that f has rational spectral factors, but that it is actually zero-free in U^2.

AD-A052 427 TEXAS TECH UNIV LUBBOCK INST FOR ELECTRONICS SCIENCE F/G 9/3
SEMIANNUAL REVIEW OF RESEARCH UNDER THE ASSOCIATE JOINT SERVICE--ETC(U)
OCT 77 R SAEKS, K S CHAO, S R LIBERTY N00014-76-C-1136
NL

UNCLASSIFIED

2 OF 2
ADA
052427

END
DATE
FILMED
5-78
DDC

Proof:

By theorem 4, it suffices to prove that f has no zeros in the set

$$\{(Z_1,Z_2) \in U^2 \mid |Z_1| = |Z_2|\}$$

But this set is the union of the open discs

$$\{(Z_1,Z_2) \mid Z_2 = e^{j\Psi}Z_1,\ |Z_1|<1\}, \text{for } 0 \leq \Psi < 2\pi;$$

we therefore wish to prove that f has no zeros in any of these discs; or equivalently that the function f_Ψ of one variable defined by $f_\Psi(Z) = f(Z,Ze^{j\Psi})$ has no zeros in the open unit disc. Applying Jensen's formula [6, p.299] for the unit disc to f_Ψ, we get

$$\frac{1}{2\pi}\int_0^{2\pi} \log \left|f_\Psi(e^{j\theta})\right| d\theta = \log \left|f_\Psi(0)\right| - \sum \log |Z_i|$$

where the summation is over all the zeros (counted with multiplicity) of f_Ψ in the unit disc. Expressing this in terms of f:

$$\frac{1}{2\pi}\int_0^{2\pi} \log \left|f(e^{j\theta},e^{j(\theta+\Psi)})\right| d\theta = \log \left|f(0,0)\right| - \sum \log |Z_i|$$

and so $\sum \log |Z_i| = 0$.

Since for any Z_i in the open unit disc $\log |Z_i| < 0$, the conclusion follows.

(It is clear from the proof that we always have

$$\frac{1}{2\pi}\int_0^{2\pi} \log \left|f(e^{j\theta},e^{j(\theta+\Psi)})\right| d\theta \geq \log \left|f(0,0)\right| \ ;$$

it follows from this that in fact the apparently weaker condition

$$\frac{1}{4\pi^2}\int_0^{2\pi}\int_0^{2\pi} \log \left|f(e^{j\theta_1},e^{j\theta_2})\right| d\theta_1 d\theta_2 = \log \left|f(0,0)\right|$$

is sufficient to guarantee that f is zero-free in U^2. See [2,p.73]).

Stable IIR Filters and Minimum-phase FIR Filters:

The very close relationship of spectral factorization to the nonvanishing of polynomials in U^2, and thereby to stable IIR filters (via the denominator polynomial) and minimum-phase filters (via the numerator polynomial) is already clear from the previous sections. The force of theorem 2 is that

purely from the point of view of amplitude response, transfer functions having rational spectral factors are equivalent to those without poles or zeros in U^2. Thus the restrictions on amplitude response in theorems 1 and 3 apply to the denominator polynomial of any stable IIR filter; the contribution of the denominator polynomial to the overall amplitude response of the filter (in the case of an all-pole filter, the entire amplitude response) must satisfy the restrictions imposed by theorems 1 and 3. We have, therefore, identified the properties of the amplitude response which make it impossible to stabilize a filter; if the original amplitude response has large overall variation in the "wrong" directions, attempting to find a stable filter which closely matches this response is futile. Close matching of the amplitude forces instability. This has already been shown by example by Bose [9] and Woods [10]: we now see that it is the variations in the amplitude response in the "wrong" directions in their examples which accounts for their behaviour.

It is also of interest to note that, in the Shanks procedure of minimizing

$$\iint |fg-1|^2 \, d\theta_1 d\theta_2$$

over all polynomials f of given degree (where g is the original polynomial), if the allowable f's were restricted to those which have polynomial spectral factorizations, the procedure would yield a polynomial devoid of zeros in U^2. It does not appear that this observation can be used as the basis for a workable stabilization method, however, since the condition that f have polynomial spectral factors is intractably nonlinear in the coefficients of f; and further, in many cases this procedure would yield an f which was only marginally stable. For the same reasons, restricting oneself throughout the design procedure to polynomials which satisfy the condition in theorem 5 does not appear to be a feasible method of ensuring stability.

Examples and Comments:

An example of the behaviour of those polynomials not possessing polynomial spectral factors has already appeared in the literature, although in a different context; we repeat this example here.

$$A(Z_1,Z_2) = 1 - .75Z_1 + .9Z_1^2 + 1.5Z_2 - 1.2Z_1Z_2 + 1.3Z_1^2Z_2 + 1.2Z_2^2 + .9Z_1Z_2^2 + .5Z_1^2Z_2^2$$

This polynomial was studied in [7]; the associated Shanks polynomial was found to be stable but to have a substantially different amplitude response from that of A (for more details, see [7]). The fact that A does not have polynomial spectral factors was established by checking the condition in theorem 3 for $m=n=1$ and $\Psi = 0$, $\Psi = \pi$, with the following results: (correct to nine decimals)

$$\frac{1}{2\pi}\int_0^{2\pi} \log \left|A(e^{j\theta}, e^{j\theta})\right| d\theta = .696570700$$

$$\frac{1}{2\pi}\int_0^{2\pi} \log \left|A(e^{j\theta}, e^{j(\theta+\pi)})\right| d\theta = 1.134686936$$

As an example of a polynomial with rational spectral factors, we have

$$B(Z_1,Z_2) = 1 + 2.25Z_1 + 2.25Z_2 + .5Z_1^2 + .5Z_2^2 - 6.5Z_1Z_2 - Z_1^2Z_2 - Z_1Z_2^2 - 4Z_1^2Z_2^2$$

This factors into $(1 + .25Z_1 + .25Z_2 + .5Z_1Z_2)(1 + 2Z_1 + 2Z_2 - 8Z_1Z_2)$ the first factor factor having no zeros in U^2, the second none in V^2; reversing the second factor gives a polynomial without zeros in U^2:

$$\tilde{B}(Z_1,Z_2) = (1 + .25Z_1 + .25Z_2 + .5\,Z_1Z_2)(-8 + 2Z_2 + 2Z_1 + Z_1Z_2)$$

$$= -8 - 2Z_1Z_2 + .5\,Z_1^2 + .5Z_2^2 + 1.25Z_1^2Z_2 + 1.25Z_1Z_2^2 + .5Z_1^2Z_2^2,$$

and $\tilde{B}$ has the same amplitude response as B.

In order to gain some idea of the stringency of the conditions in theorem 3, let us consider the case of an ideal band-pass filter. By an ideal band-pass filter we will mean a filter whose amplitude response is equal to 1 on some subset, A, of the square $o \leq \theta_1 < 2\pi$, $o \leq \theta_2 < 2\pi$, and equal to $K << 1$ on the complement of A (of course this specification continues over the whole plane by periodicity). This of course is not the amplitude response of any rational function, but in practice for certain shapes of the set A, one may wish to approximate such a response by a rational function. One easily sees that up to a scale factor, the averages in theorem 3 are in this case merely the fraction:

$$\frac{\text{length of the line } L_i \text{ lying in the complement of A}}{\text{Total length of the line } L_i}$$

It is easily seen from this that there are very few passband shapes of practical interest which satisfy even the first of these conditions (where n=1 and m=1); in other words, very few which can be accurately approximated by transfer functions having rational spectral factors. (This is not to imply that one would in practice be restricted to such filters; the above discussion is meant solely as an indication of the severity of the restrictions on the amplitude of such filters).

Finally, we remark that there does not seem to be any difficulty in extending the results in this paper to higher dimensions, and to multidimensional systems other than digital filters.

APPENDIX

The converses to Theorems 1 and 3.

These converses involve some technical ideas and results from [2]; the most important ideas are those of inner function [2,p.105], outer function [2,p. 72], Poisson integral [2, p.17] and the classes $N(U^2)$ [2, p.44] and $N_*(U^2)$ [2, p.44].

We will also use the following notation from [2] (Here f is an analytic function on U):

i. $f^*(e^{j\theta_1}, e^{j\theta_2}) \triangleq \lim_{r \to 1^-} f(re^{j\theta_1}, re^{j\theta_2})$
will denote the radial limit of f
(this is clearly consistent with our previous use of f^*);

ii. For $w = (w_1, w_2) \in T^2$, $f_w(Z)$ will denote the one-varible function defined by

$$f_w(Z) \triangleq f(Zw_1, Zw_2);$$

iii. if ϕ is a function defined on T^2 which is absolutely integrable there,

$$\hat{\phi}(m,n) \triangleq \frac{1}{4\pi^2} \int_0^{2\pi} \int_0^{2\pi} \exp(-jm\theta_1 - jn\theta_2)\, \phi(\theta_1,\theta_2)\, d\theta_2 d\theta_1$$

will denote the Fourier coefficients of ϕ.

iv. For any function ϕ on T^2,

$\frac{1}{4\pi^2} \int_0^{2\pi} \int_0^{2\pi} \phi(\theta_1,\theta_2)\, d\theta_2 d\theta_1$ will be denoted by

$$\int_{T^2} \phi \, dm \quad \text{or} \quad \int_{T^2} \phi(w)\, d\, m\, (w).$$

We will first prove the converse to Theorem 1, and from this derive the converse to Theorem 3. First of all, however, we need the following

lemma (which is given as a problem in [2]).

Lemma A1:

If ϕ is a real-valued function defined on T^2 such that

$$\phi \in L^1(T^2) \quad (\text{i.e., } \int_{T^2} |\phi| dm < \infty)$$

and

$$\hat{\phi}(m,n) = 0 \quad \text{for } mn < 0,$$

then there is an outer function f on U^2 such that

$$P[\phi] = \log|f|$$

(where P[ ] denotes "Poisson integral of").

Proof

$$\text{Let } a_{mn} = \begin{cases} \hat{\phi}(m,n) & (m,n) \neq (o,o) \\ 1/2\ \hat{\phi}(m,n) & (m,n) = (o,o) \end{cases}$$

and let

$$g(Z_1,Z_2) = \sum_{m=o}^{\infty} \sum_{n=o}^{\infty} a_{mn} Z_1^m Z_2^n.$$

This series clearly converges uniformly on compact subsets of U^2, and so defines an analytic function there.

If we let $f = e^g$

then f is analytic in U^2, and

$$\log |f| = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} a_{mn} r_1^m r_2^n \exp(jm\theta_1 + jn\theta_2)$$

$$+ \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \overline{a_{mn}} r_1^m r_2^n \exp(-jm\theta_1 - jn\theta_2)$$

$$= \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \hat{\phi}(m,n) r_1^{|m|} r_2^{|n|} \exp(jm\theta_1 + jn\theta_2)$$

$$= P[\phi] \qquad [2,,p.17]$$

Next we prove that f is outer; we have (for $0<r<1$)

$$\int_{T^2} \log^+ |f(rw)| dm(w) \leq \int_{T^2} |\log |f(r,w)|| dm(w)$$

$$= \int_{T^2} |P[\phi](rw)| dm(w)$$

$$\leq \int_{T^2} |\phi(w)| dm(w) \qquad [2,\text{Thm.2.1.3(c)}]$$

$$< \infty$$

and so $f \in N(U^2)$.

Now f^* exists almost everywhere on T^2 [2,Thm.3.3.5] and $\log|f^*| = \phi$ almost everywhere on T^2 [2,Thm.2.2.1]; thus $\log|f| = P[\log|f^*|]$ and so $f \in N_*(U^2)$ [2,Thm.3.3.5], and $\log|f(0)| = \int_{T^2} \log|f^*(w)| dm(w)$.

Thus f is outer.

Q.E.D.

We can now prove the converse to Theorem 1:

Theorem A2:

Let $f(Z_1,Z_2)$ be a rational function ($\not\equiv 0$), and let

$$\phi = \log |f^*|$$

If $\hat{\phi}(m,n) = 0$ for $mn<0$, then there is a rational function g without poles or zeros in U^2 such that $|g^*| = |f^*|$.

Proof:

By Lemma A1, there is an outer function g such that

$$\log |g| = P[\log|f|].$$

This implies

$$\log |g^*| = \log |f^*| \quad \text{almost everywhere on } T^2.$$

Therefore, for almost all $w \in T^2$

$$\log |g_w^*(Z)| = \log |f_w^*(Z)| \quad \text{for almost all } Z \in T \quad [2,\text{Lemma 3.3.2}],$$

and g_w is outer for almost all $w \in T^2$ [2,Lemma 4.4.4].

For any such w, let $Z_1, \ldots, Z_n$ denote the poles, and $Z_{n+1}, \ldots Z_m$ the zeros, of $f_w(Z)$ in U, and let

$$\tilde{f}_w(Z) = \prod_{k=1}^{n} \frac{Z-Z_k}{\bar{Z}_k Z-1} \prod_{k=n+1}^{m} \frac{\bar{Z}_k\, Z-1}{Z-Z_k} \; f_w(Z)$$

Then $\tilde{f}_w$ has no poles or zeros in U and is rational; hence, $\tilde{f}_w$ is outer. Since g_w is outer, we have $\tilde{f}_w/g_w$ is outer. Also $|\tilde{f}_w^*| = |f_w^*|$, and so $|\tilde{f}_w^*| = |g_w^*|$ for almost all $Z \in T$. Thus $\tilde{f}_w/g_w$ is inner. But a function which is both outer and inner is a constant of modulus 1, and so

$$g_w = e^{j\Psi} \tilde{f}_w \qquad \text{for some real } \Psi.$$

Thus g_w is rational for almost all $w \in T^2$, and so g_w is rational for all $w \in E$, where $E \subseteq T^2$ is a compact set of positive measure (by the inner regularity of the measure). *It follows by [2,Thm. 5.2.2] that g is* rational (since the vanishing of a polynomial P on a set of positive measure in T^2 would imply

$$\log |P^*| \notin L^1(T^2)$$

and so $P \equiv 0$.)

Thus g is a rational function without poles or zeros in U^2, and

$$|g^*| = |f^*| \quad \text{almost everywhere in } T^2$$

and so, since g and f are both rational,

$$|g^*| = |f^*| \quad \text{on } T^2.$$

Q.E.D.

We next prove the converse to Theorem 3:

Theorem A3:

Let $f(Z_1,Z_2)$ be a rational function ($\not\equiv 0$) and let

$$\Phi = \log |f^*| .$$

If $\frac{1}{2\pi} \int_0^{2\pi} \Phi(m\theta, n\theta + \Psi)d\theta$ is a constant independent of Ψ for each pair (m,n) with $m > o$ and $n > o$ then there is a rational function g without poles or zeros in U^2 such that $|g^*| = |f^*|$.

Proof:

Let $m > o$, $n > o$, and let $\ell \neq o$ be an integer.

Then

$$\int_0^{2\pi} e^{j\ell m\Psi} \int_0^{2\pi} \Phi(m\theta, n\theta + \Psi) d\theta \, d\Psi = 0$$

$$\Rightarrow \quad \int_0^{2\pi} \int_0^{2\pi} e^{j\ell m\Psi} \; \Phi\,(m\theta, n\theta + \Psi) d\theta d\Psi = 0$$

Making the change of variables defined by

$$\theta = \frac{1}{m}\theta_1$$
$$\Psi = \theta_2 - \frac{n}{m}\theta_1,$$

we get

$$\frac{1}{m} \int_0^{2\pi} \int_{\frac{n}{m}\theta_1}^{\frac{n}{m}\theta_1 + 2\pi} \exp\,(j\ell m\theta_2 - j\ell n\theta_1)\; \Phi\,(\theta_1, \theta_2) d\theta_2 d\theta_1 = 0.$$

and since the integrand is periodic in θ_1 and θ_2

$$\int_0^{2\pi} \int_0^{2\pi} \exp\,(j\ell m\theta_2 - j\ell n\theta_1)\; \Phi\,(\theta_1, \theta_2) d\theta_2 d\theta_1 = 0$$

and so $\hat{\Phi}\,(-\ell n, \ell m) = 0$ for all $\ell \neq 0$, $m > o$ and $n > 0$,

that is,

$$\hat{\Phi}(m,n) = 0 \quad \text{for all } m,n \text{ with } mn < o.$$

The result now follows from Theorem A2.

Q.E.D.

Finally, we note that if f in Theorem A3 is a polynomial, then the converse in Theorem 2 implies that f has polynomial spectral factors. Thus we have the full converse of Theorem 3 for polynomials.

References

1. M.P. Ekstrom, and J.W. Woods, "Two-Dimensional Spectral Factorization with Application to Recursive Digital Filtering", IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-24, pp. 115-128, April 1976.

2. W. Rudin, Function Theory in Polydiscs, New York, Benjamin, 1969.

3. W. Stoll, Holomorphic Functions of Finite Order in Several Complex Variables, (CBMS Regional Conference Series in Mathematics) Providence, R.I., AMS, 1973.

4. R.A. DeCarlo, J. Murray, and R. Saeks, "Multivariable Nyquist Theory", International Journal of Control, (to appear).

5. T.S. Huang, "Stability of Two-Dimensional Recursive Filters", IEEE Trans. Audio Electroacoust., Vol. AU-20, pp. 158-163, June 1972.

6. W. Rudin, Real and Complex Analysis, New York, McGraw-Hill, 1966.

7. J.L. Shanks, S. Treitel, and J.H. Justice, "Stability and Synthesis of Two-Dimensional Recursive Filters", IEEE Trans. Audio Electroacoust, Vol. AU-20, pp. 115-128, June 1972.

8. R.C. Gunning, and H. Rossi, Analytic Functions of Several Complex Variables, Englewood Cliffs, NJ, Prentic-Hall, 1965.

9. N.K. Bose, "Problems in Stabilization of Multidimensional Filters via Hilbert Transform", IEEE Trans. Geo. Sci. Elec., Vol. GE-12, pp. 146-147, October 1974.

10. J.W. Woods, Correspondence in IEEE Trans. Geo.Sci. Elec., Vol. GE-12, p. 104, July 1974.

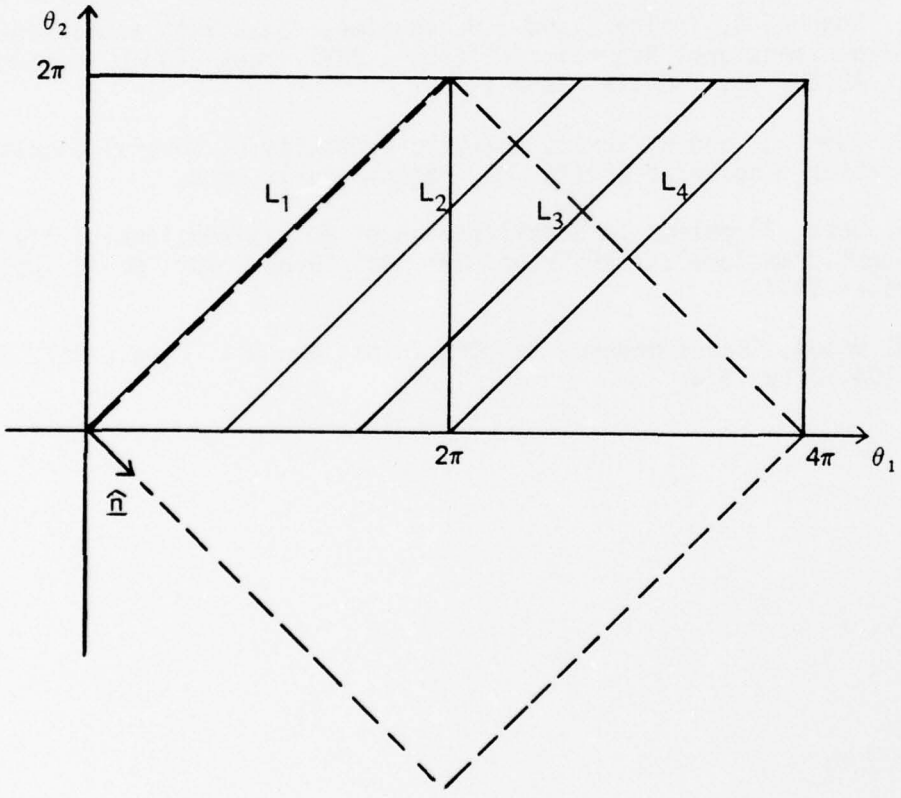


Figure 1.

RESEARCH

on

OPTICAL NOISE

G. Froehlich and J. Walkup

DEPARTMENT OF ELECTRICAL ENGINEERING

TEXAS TECH UNIVERSITY

ABSTRACT

Optimal estimators are derived for a class of signal-dependent noise processes. Such processes are of interest in optics because certain phenomena, such as film grain strates that when one ignores the presence of signal-dependent noise and instead assumes only signal-independent noise models, the resulting estimators may pay a severe penalty in performance. This "mismatch" problem is explored, with the results of Monte Carlo simulations of the performances of both optimum and mismatched estimators being presented. The Cramer-Rao lower bounds on the mean square estimation errors for unbiased estimators are evaluated and compared with the lower bounds derived for the signal-independent noise case. Overall, the results indicate that improved performance will, in most cases, offset the increased complexity inherent in estimators designed for the signal-dependent noise model.

Introduction

In contrast to the signal-independent additive noise models traditionally encountered in statistical communication theory [1,2], many physical noise processes are inherently signal-dependent. Common examples from optical processing include film grain noise, encountered in image processing, and photoelectronic shot noise, which is sometimes dominant when imaging at low light levels with photoemissive detectors.[3,4] An example of a non-optical noise source which is effectively signal-dependent is magnetic tape recording noise.[5] A study of these particular examples indicates that studies of optimum estimation in signal-dependent noise processes would have applications to a broad class of signal processing problems in modern optics and in other fields.

To date, the majority of the work dealing with signal-dependent noise has been concentrated on rather specialized examples and applications. Using a Poisson point process noise model, Goodman and Belsher [6] have considered the restoration of atmospherically degraded images using linear minimum mean square error filters. Walkup and Choens[4] modified the familiar Wiener filter for various additive, Gaussian signal-dependent noise models, and Naderi[7] has done considerable additional work on this problem. Additionally, Hunt[8] has derived a nonlinear maximum a posteriori (MAP) estimator, based on a different model than the one considered here, which can accomodate both signal-dependent and signal-independent noise cases, and they have applied this MAP estimator to restoring noise-degraded images. For such applications, and in the special case where the images of interest exhibit extremely low contrasts, conventional restoration techniques perform rather poorly. Thus, heuristic algorithms, such as the so-called "noise cheating" algorithm for film grain noise suppression[9], have been developed. Other algorithms, which explicitly include the signal dependence of the noise, as well as incorporating pertinent properties of the human visual system, have also been investigated[7,10,11].

The purpose of this paper, then, is twofold. First, several fundamental properties of signal-dependent noise are investigated in order to better understand when consideration of signal-dependence is warranted and when it can be ignored. To this end, the mean square estimation error is first considered for both the signal-dependent and signal-independent cases. In addition, the mean square estimation error for a

mismatched case is evaluated. The mismatch case considered is one in which the signal-dependent measurement model is valid but is ignored for purposes of simplification. Secondly, optimal estimators are derived for several cases of both signal-dependent and signal-independent models. The Cramér-Rao lower bound on mean square estimation error is also determined, in order to find the lowest error possible for both signal-dependent and signal-independent estimators. The results of Monte Carlo simulations of the performance of the various optimal estimators previously derived are presented for several values of the model parameters and for various prior signal probability densities.

Problem Statement

To motivate the investigation of signal-dependent noise processes, it is necessary first to define the models to be used. The signal-dependent measurement model to be used is given by

$$r = s + kf(s)n_1 + n_2, \qquad (1)$$

where n_1 and n_2 are signal-independent random noise processes; s is the underlying signal to be estimated which is assumed to have probability density p(s); n_1, n_2, and s are assumed mutually statistically independent; f(s) is any function of the signal; k is a scalar constant; and r is the noisy measurement. The signal-dependent noise term in Eq. (1) is, of course, the term $kf(s)n_1$. It is often physically reasonable to assume that both n_1 and n_2 are zero mean and have unimodal probability densities. Further, note that substitution of k = 0 in Eq. (1) yields

$$r = s + n_2, \qquad (2)$$

which is just the familiar textbook additive, signal-independent noise model[1,2]. In both Eq. (1) and Eq. (2), the arguments of all of the variables have been dropped for simplification. It should be remembered that these arguments may depend on time, position, or both.

It will be shown repeatedly that the model of Eq. (2) yields far simplier estimators than does Eq. (1), as would be expected. The following example serves to illustrate why it may prove worthwhile to employ the more complex estimators resulting from Eq. (1).

When the observations are actually of the type given by Eq. (2), it can be shown[2] that simply using the received value as the estimate results in a minimum-variance unbiased estimate, i.e.,

$$\hat{s} = r, \tag{3}$$

where the circumflex denotes the estimate. The average error is then given by

$$E\{\hat{s} - s\} = E\{r - s\} = E\{n_2\} = 0. \tag{4}$$

The estimator is said to be unbiased since the mean error is zero. A measure of the performance of this estimator, conditioned on the signal value, is given by the conditional mean square error, and is found to be

$$E\{(\hat{s} - s)^2 | s\} = E\{n_2^2\} = \sigma_2^2, \tag{5}$$

which is simply the variance of the additive noise process n_2. This estimator is obviously simple from an implementation point of view.

With this in mind, consider a case in which the observations are actually of the type given by the signal-dependent model of Eq. (1). For ease of implementation it is decided to use the estimate given in Eq. (3), which was designed for the signal-independent noise process. This represents a mismatched situation, where an estimator based upon an incorrect measurement model (corresponding to ignoring the signal-dependency) is used. Once again, the average estimation error is zero, due to n_1 and n_2 being assumed zero mean and to the assumed mutual statistical independence of n_1, n_2, and s.

However, assuming $\hat{s} = r$, the mean square estimation error for this mismatched case is given by

$$E\{(\hat{s} - s)^2 | s\} = k^2\sigma_1^2 \, E\{[f(s)]^2\} + \sigma_2^2 \, , \quad (6)$$

For convex f(s), i.e. $f''(s) \geq 0$ for all s, Jensen's inequality states that $E\{f(s)\} \geq f[E\{s\}]$[12], where $E\{\cdot\}$ denotes the expected value. This inequality may be used to find a lower bound for the mean square estimation error for the mismatched case. Thus, recalling Eqs. (5) and (6),

$$\sigma_2^2 \leq k^2\sigma_1^2\{f[E(s)]\}^2 + \sigma_2^2 \leq k^2\sigma_1^2 \, E\{[f(s)]^2\} + \sigma_2^2. \quad (7)$$

Note that this gives a lower bound (the middle term) on the mean square estimation error of the mismatched estimator, and that this bound contains a function of the signal's mean. The leftmost term of Eq. (7) is the mean square estimation error given by Eq. (5). Note that the mismatched mean square estimation error is in general greater than the error for the same estimator when used in the presence of signal-independent noise. We next consider an illustration of the significance of Eq. (7).

A commonly used model in image processing when the observed quantity is the photographic density is given by Eq. (1) with f(s) given by s^p.[4,10] From Eq. (6), then, the mismatched mean square estimation error becomes

$$E\{(\hat{s} - s)^2|s\} = k^2 \sigma_1^2 E\{s^{2p}\} + \sigma_2^2 , \quad (8)$$

where k is a scanning constant relating the scanning aperture area to the mean area of a film grain. A typical value of p used for characterizing photographic film-grain noise is $p = 1/2$, though $p = 1/3$ has also been used.[4,10] Thus, Eq. (8) becomes (for $p = 1/2$)

$$E\{(\hat{s} - s)^2|s\} = k^2\sigma_1^2 E\{s\} + \sigma_2^2, \quad (9)$$

which is greater than the variance of Eq. (5) by the addition of a term which is proportional to the signal mean. Note that in the particular case of $p = 1/2$, the equality holds

between the last two terms in Eq. (7), but that for general p this is not the case. Here, the lower bound on the mean square estimation error given by Eq. (7) becomes

$$E\{(\hat{s} - s)^2 | s\} \geq k^2\sigma_1^2[E(s)]^{2p} + \sigma_2^2. \qquad (10)$$

The lower bound given by Eq. (10) may be visualized with the aid of Figs. 1 and 2, for various values of k, p, and E(s). In all cases, the plane upon which the surfaces rest is not the zero plane, but rather represents a height of σ_2^2, the leftmost term of Eq. (7), which results when the estimator of Eq. (3) is properly matched (to the signal-independent noise process of Eq. (2)). In Fig. 1, p is fixed at a value of 1/2, σ_1^2 and σ_2^2 are set equal to 1 for illustration, with k and E(s) being varied. In Fig. 2, k is fixed (at k = 1/2) and p is varied. It should be noted that, for film grain noise applications, common values of k are in the range of from about 0.3 to about 0.7. These figures illustrate the marked deviation from the variance achieved by a properly matched signal-independent estimator. Also, it should be remembered that these surfaces represent lower bounds on the mean square estimation error of the mismatched estimator, and there is no guarantee, in general, that even this measure of performance can be achieved. Thus, optimal estimators based on the proper noise model are needed. These estimators are derived in the following sections.

MAP Estimation

An appropriate optimal estimate when the signal is random and its probability density function is known *a priori* is the maximum a posteriori probability (MAP) estimate.[2] This estimate, $\hat{s}_{MAP}$, is defined to be that value of s which maximizes the *a posteriori* density $p(s|r)$. In other words, given the observation r, the signal value $\hat{s}_{MAP}$ maximizes the probability of that value of r having been received. Maximizing $p(s|r)$ is equivalent to maximizing $p(r|s)p(s)$, or alternately the logarithm of this product. This follows from the facts that (a)

$$p(s|r) = \frac{p(r|s)p(s)}{p(r)} , \qquad (11)$$

(b) the denominator is not a function of s, and (c) because monotonic transformations (such as the logarithm) preserve maxima and minima.

As an example of the calculation of a MAP estimate, assume that n_1 and n_2 are both zero mean, normally distributed random variables having variances σ_1^2 and σ_2^2, respectively. In this case, the conditional probability density $p(r|s)$ is also normal, with a mean of s and variance v(s), given by

$$v(s) = k^2\sigma_1^2[f(s)]^2 + \sigma_2^2. \qquad (12)$$

It can then be shown that the MAP estimate is a solution of the equation

$$(r - \hat{s}_{MAP})^2 v'(\hat{s}_{MAP}) + 2(r - \hat{s}_{MAP})v(\hat{s}_{MAP}) - v'(\hat{s}_{MAP})v(\hat{s}_{MAP})$$

$$+ 2[v(\hat{s}_{MAP})]^2 \frac{\partial}{\partial \hat{s}_{MAP}} \ln p(\hat{s}_{MAP}) = 0, \qquad (13)$$

where the prime denotes the partial derivative with respect to $\hat{s}_{MAP}$.

For the class of situations where $f(s) = s^p$, and assuming s is distributed normally with mean μ_s and variance σ_s^2, Eq. (13), the MAP equation becomes

$$[2k^2\sigma_1^2(p-1) - \frac{4k^2\sigma_1^2\sigma_2^2}{\sigma_s^2}] \hat{s}_{MAP}^{2p+1} + [2rk^2\sigma_1^2(1-2p) + \frac{4k^2\sigma_1^2\sigma_2^2\mu_s}{\sigma_s^2}] \hat{s}_{MAP}^{2p}$$

$$- [2\sigma_2^2 + \frac{2\sigma_2^4}{\sigma_s^2}] \hat{s}_{MAP} + [2pk^2\sigma_1^2(r^2-\sigma_2^2)] \hat{s}_{MAP}^{2p-1}$$

$$+ [2\sigma_2^2 r + \frac{2\sigma_2^4\mu_s}{\sigma_s^2}] - [2pk^4\sigma_1^4] \hat{s}_{MAP}^{4p-1} + [\frac{2k^4\sigma_1^4\mu_s}{\sigma_s^2}] \hat{s}_{MAP}^{4p}$$

$$- [\frac{2k^4\sigma_1^4}{\sigma_s^2}] \hat{s}_{MAP}^{4p+1} = 0. \qquad (14)$$

The MAP estimate, $\hat{s}_{MAP}$, is a solution of Eq. (14). For the specific case where $p = 1/2$, Eq. (14) reduces to the cubic equation

$$[\frac{2k^4\sigma_1^4}{\sigma_s^2}] \hat{s}_{MAP}^3 + [\frac{4k^2\sigma_1^2\sigma_2^2 - 2k^4\sigma_1^4\mu_s}{\sigma_s^2} + 2k^2\sigma_1^2] \hat{s}_{MAP}^2$$

$$+ [\frac{2\sigma_2^4 - 4k^2\sigma_1^2\sigma_2^2\mu_s}{\sigma_s^2} + k^4\sigma_1^4 + 2\sigma_2^2] \hat{s}_{MAP}$$

$$+ [k^2\sigma_1^2(\sigma_2^2 - r^2) - 2\sigma_2^2 r - \frac{2\sigma_2^4\mu_s}{\sigma_s^2}] = 0. \qquad (15)$$

Substitution of k = 0 into Eq. (14) or Eq. (15) yields the MAP estimate for the signal-independent noise case of Eq. (2), namely

$$\hat{s}_{MAP} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_2^2} r + \frac{\sigma_2^2}{\sigma_s^2 + \sigma_2^2} \mu_s. \tag{16}$$

Comparison of Eqs. (14) and (16) demonstrates the much greater complexity of the estimator structure when signal-dependent noise processes are taken into account.

A shortcoming of the Gaussian density as a model for p(s) is that the probability of a negative value of s is nonzero, regardless of E(s), and often this can be physically impossible (for example, when s represents photographic density). However, one common probability density of interest for some classes of images is the Rayleigh density,[13] i.e.,

$$p(s) = \frac{s}{\sigma^2} \exp[-\frac{s^2}{2\sigma^2}], \; s \geq 0, \tag{17}$$

which has a positivity constraint.

Substitution into Eq. (13), the MAP equation, with $f(s) = s^p$ as before, yields

$$[2k^2\sigma_1^2(p-1) - \frac{4k^2\sigma_1^2\sigma_2^2}{\sigma^2}]\hat{s}_{MAP}^{2p+2} + [2k^2\sigma_1^2 r(1-2p)]\hat{s}_{MAP}^{2p+1}$$

$$-[2\sigma_2^2 + \frac{2\sigma_2^4}{\sigma^2}]\hat{s}_{MAP}^2 + [2pk^2\sigma_1^2(r^2-\sigma_2^2 + \frac{2\sigma_2^2}{p})]\hat{s}_{MAP}^{2p}$$

$$+[2\sigma_2^2 r]\hat{s}_{MAP} - [2k^4\sigma_1^4(p-1)]\hat{s}_{MAP}^{4p} + [2\sigma_2^4]$$

$$-[\frac{2k^4\sigma_1^4}{\sigma^2}]\hat{s}_{MAP}^{4p+2} = 0. \tag{18}$$

This equation is quite similar to Eq. (14) with $\mu_s = 0$, but each term is greater in Eq. (18) by degree one. Thus, when $p = 1/2$, the MAP estimate $\hat{s}_{MAP}$ is a solution of the quartic

$$[\frac{2k^4\sigma_1^4}{\sigma^2}]\hat{s}_{MAP}^4 + [k^2\sigma_1^2(1 + \frac{4\sigma_2^2}{\sigma^2})]\hat{s}_{MAP}^3 + [2\sigma_2^2 + \frac{2\sigma_2^4}{\sigma^2} - k^4\sigma_1^4]\hat{s}_{MAP}^2$$

$$-[k^2\sigma_1^2(r^2 + 3\sigma_2^2) + 2r\sigma_2^2]\hat{s}_{MAP} - 2\sigma_2^4 = 0. \qquad (19)$$

As before, substitution of $k = 0$ into Eq. (19) then yields the MAP estimate for the signal-independent noise case, given by

$$\hat{s}_{MAP} = \frac{\sigma^2}{\sigma^2 + \sigma_2^2} r + \frac{[4\sigma^2\sigma_2^2(\sigma^2 + \sigma_2^2) + r^2\sigma^4]^{1/2}}{2(\sigma^2 + \sigma_2^2)} . \qquad (20)$$

Again this is less complex than the MAP equation of Eq. (18), but it is not as simple as the solution for the signal-independent noise model with a normal probability density for s.

Another probability density with a positivity constraint is the folded normal, that is, the absolute value of a normally distributed random variable. Its probability density function is given by

$$p(s) = [p_N(s) + p_N(-s)]u(s), \qquad (21)$$

where $u(s)$ is the unit step function and $p_N(s)$ is the normal probability density function. After much manipulation, it may be shown that the MAP estimate for this case is given by the value of $\hat{s}_{MAP}$ which satisfies

$$\exp\left[\frac{2\mu_s \hat{s}_{MAP}}{\sigma_s^2}\right]\left\{\frac{\mu_s - \hat{s}_{MAP}}{\sigma_s^2} + \frac{r - \hat{s}_{MAP}}{2v(\hat{s}_{MAP})}\left[1 + \frac{(r-\hat{s}_{MAP})v'(\hat{s}_{MAP})}{v(\hat{s}_{MAP})}\right]\right\}$$

$$-\left\{\frac{\mu_s + \hat{s}_{MAP}}{\sigma_s^2} - \frac{r - \hat{s}_{MAP}}{2v(\hat{s}_{MAP})}\left[1 + \frac{(r-\hat{s}_{MAP})v'(\hat{s}_{MAP})}{v(\hat{s}_{MAP})}\right]\right\} = 0, \tag{22}$$

where $v(s)$ is given by Eq. (12), and the prime again denotes differentiation with respect to s. To obtain the MAP estimate for the signal-independent measurement model of Eq. (2), $k=0$ is substituted into Eq. (22) to obtain

$$\exp\left(\frac{2\mu_s \hat{s}_{MAP}}{\sigma_s^2}\right)\left[\frac{\mu_s - \hat{s}_{MAP}}{\sigma_s^2} + \frac{r-\hat{s}_{MAP}}{2\sigma_2^2}\right] - \left[\frac{\mu_s + \hat{s}_{MAP}}{\sigma_s^2} - \frac{r-\hat{s}_{MAP}}{2\sigma_2^2}\right] = 0. \tag{23}$$

Neither of these equations lend themselves to straightforward solution: however, it is once again obvious that the signal-independent noise model yields a much simpler solution.

ML Estimation

Another commonly used estimator is the maximum likelihood (ML) estimator.[2] The ML estimate is employed when no prior knowledge of the signal is assumed, and it is found by maximizing $p(r|s)$ over s. In other words find a value of s, such that given s, the most probable observation r which would result is the value observed. Using the signal-dependent measurement model of Eq. (2), and still assuming n_1 and n_2 are zero mean normal random variables with variances σ_1^2 and

σ_2^2, respectively, the ML estimate, $\hat{s}_{ML}$, is a solution of the equation

$$(r-\hat{s}_{ML})^2 v'(\hat{s}_{ML}) + 2(r-\hat{s}_{ML})v(\hat{s}_{ML}) - v'(\hat{s}_{ML})v(\hat{s}_{ML}) = 0, \tag{23}$$

where $v(\hat{s}_{ML})$ and $v'(\hat{s}_{ML})$ are as defined previously. Again, considering the special case $f(s) = s^p$, the ML equation becomes

$$[2k^2\sigma_1^2(p-1)]\hat{s}_{ML}^{2p+1} + [2rk^2\sigma_1^2(1-2p)]\hat{s}_{ML}^{2p} + [2pk^2\sigma_1^2(r^2-\sigma_2^2)]\hat{s}_{ML}^{2p-1}$$

$$-[2\sigma_2^2]\hat{s}_{ML} - [2pk^4\sigma_1^4]\hat{s}_{ML}^{4p-1} + 2\sigma_2^2 r = 0. \tag{24}$$

This equation is at worst no more complex than the MAP equation, Eq. (14). In fact, for $p = 1/2$, Eq. (24) becomes the quadratic equation

$$(k^2\sigma_1^2)\hat{s}_{ML}^2 + [2\sigma_2^2 + k^4\sigma_1^4]\hat{s}_{ML} + [k^2\sigma_1^2(\sigma_2^2 - r^2) - 2\sigma_2^2 r] = 0. \tag{25}$$

This has as its positive root the ML estimate

$$\hat{s}_{ML} = \left[r^2 + \left(\frac{k^2\sigma_1^2}{2}\right)^2 + \frac{2r\sigma_2^2}{k^2\sigma_1^2} + \left(\frac{\sigma_2^2}{k^2\sigma_1^2}\right)^2\right]^{1/2} - \frac{k^2\sigma_1^2}{2} - \frac{\sigma_2^2}{k^2\sigma_1^2}. \tag{26}$$

The ML estimate for the signal-independent model of Eq. (2) is found by letting $k = 0$ in any of Eqs. (23) to (25), and is given by

$$\hat{s}_{ML} = r. \tag{27}$$

Note that this is the minimum variance unbiased estimate used in Eq. (3) for the mismatched example, for which we earlier found the mean square estimation error.

Another point worthy of note is the similarity between Eq. (13), the general MAP equation, and the ML equation, Eq. (23). These expressions differ only by an additional term in Eq. (13), and it is this term which contains all of the prior knowledge about s. This term vanishes when $\ell np(s)$, and hence p(s), is constant. In other words, if s is distributed uniformly over all of its space of definition (a worst case), then knowledge of its value in no way affects the maximum of $p(s|r) = p(r|s)p(s) \doteq c\, p(r|s)$. Thus, the ML estimator can be viewed as a worst case of the MAP estimator. Because the MAP estimator embodies a priori information about s that is not present in the formulation of the ML estimate, it would seem reasonable to assume that the MAP estimate would exhibit a smaller mean square estimation error than the ML estimate. It will be seen that this is indeed the case. In the next section, bounds on the variances of these estimates will be found.

Cramér-Rao Lower Bounds

A well-known lower bound on the variance of any unbiased estimate for a fixed but unknown s is the Cramér-Rao error

bound.[2] Given the conditional density $p(r|s)$, the Cramér-Rao bound is given by

$$\text{var}[\hat{s}-s|s] \geq \{-E[\frac{\partial^2 \ell np(r|s)}{\partial s^2}]\}^{-1} . \tag{28}$$

For n_1 and n_2 normal with zero mean, Eq. (28) reduces to

$$\text{var}[\hat{s}-s|s] \geq \frac{2[v(s)]^2}{2v(s)+[v'(s)]^2}, \tag{29}$$

where $v(s)$ and $v'(s)$ are as given by Eq. (12). For the signal-independent noise model, which is the result of letting $k = 0$ in Eq. (29), the Cramér-Rao bound is given by

$$\text{var}[\hat{s} - s|s] \geq \sigma_2^2, \tag{30}$$

which is the variance actually achieved by the ML estimate of Eq. (27) for the signal-independent noise case. When equality holds in Eq. (28), the estimate $\hat{s}$ is said to be efficient (2). Thus, the signal-independent ML estimate is efficient when the measurement is actually of the form given by Eq. (1).

For $f(s) = s^p$, as before, Eq. (29) becomes

$$\text{var}[\hat{s}-s|s] \geq \frac{k^4\sigma_1^4 s^{4p}+2k^2\sigma_1^2\sigma_2^2 s^{2p}+\sigma_2^4}{k^2\sigma_1^2 s^{2p}+2p^2k^4\sigma_1^4 s^{4p-1}+\sigma_2^2} , \tag{31}$$

which for $p = 1/2$ reduces to

$$\text{var}[\hat{s}-s|s] \geq \frac{k^4\sigma_1^4 s^2+2k^2\sigma_1^2\sigma_2^2 s + \sigma_2^4}{(k^2\sigma_1^2+1/2k^4\sigma_1^4)s + \sigma_2^2} . \tag{32}$$

Although it is not obvious by inspection, the bound given by Eq. (31) may actually be smaller than the bound given by Eq. (30). In other words, there are potentially cases where the estimators designed for the signal-dependent measurement model may actually outperform (in a mean square estimation error sense) the estimators designed for the signal-independent measurement model. To better illustrate this, Eq. (31) is plotted in Figs. 3 and 4. In the first of these k is fixed at 1/2, σ_1^2 and σ_2^2 at one, and s and p are varied. As in Figs. 1 and 2, the plane upon which the surface rests is not the zero plane, but rather is the Cramér-Rao lower bound given by Eq. (30), namely σ_2^2. In Fig. 4, p is fixed at 1/2 and $k\sigma_1$ is allowed to vary. Now it is worth noting that in all of the previous equations, when $k \neq 0$, k and σ_1 always appear together. Thus varying $k\sigma_1$ is tantamount to fixing either one and varying the other. Note that in Fig. 4, for certain values of $k\sigma_1$ and s, the Cramér-Rao bound of Eq. (31) dips below the Cramér-Rao bound of Eq. (30), that is, it dips below the plane σ_2^2. This is, of course, the region mentioned above, where the inclusion of signal-dependence in the measurement model may potentially result in improved estimator performance. The values of s and $k\sigma_1$ which result in this region are given by

$$0 \geq s \leq \frac{\sigma_2^2}{2}\left[1 - \frac{2}{(k\sigma_1)^2}\right], \qquad (33)$$

where $k\sigma_1$ must then satisfy

$$k\sigma_1 \geq \sqrt{2} \quad . \tag{34}$$

Recall that these equations are derived for the p = 1/2 case.

To get a feeling for the actual mean square estimation error achieved by the estimators derived above, Monte Carlo simulations were performed, with the results presented in the next section.

Monte Carlo Simulations

The performance of each of the estimators derived in the previous sections was evaluated by Monte Carlo simulations to determine the mean square estimation error. The results for each of the various signal probability densities were so similar that only one case is presented. The Gaussian case was chosen since, for the MAP estimate, it represents the minimum achievable mean square estimation error (see Appendix). Figure 5 shows the mean square estimation error (MSEE) of the MAP estimate plotted as a function of the signal mean E(s). In Fig. 5a, $k\sigma_1 = 1$, while in Fig. 5b, $k\sigma_1 = 2$. The solid line is the MSEE for the MAP estimator of Eq. (15) and the dashed line is the MSEE for the mismatched case, that is, for the MAP estimate of Eq. (16) when applied to the signal-dependent measurement. Inclusion of signal-dependence in

the estimator structure is seen to yield estimates of the signal which, on the average, have smaller error than would be the case when signal-dependence is ignored. It should be noted that for sufficiently small $k\sigma_1$ and small signal means the signal-dependent noise term is negligible. This results in the estimates for the mismatched case being very nearly equal to those which include the signal-dependence.

Figure 6 presents the results of simulations of the ML estimators. As before, the solid line represents the signal-dependent estimator MSEE and the dashed line represents the MSEE for the mismatched case. Once again, inclusion of signal-dependence is seen to yield better estimates on the average. Since the ML estimates include no prior knowledge of the signal statistics, their performance is markedly inferior to the MAP estimates, but as previously discussed, the ML estimate represents a worst case. As before, for small $k\sigma_1$ and small E(s), the estimates are very nearly equal regardless of the inclusion of signal-dependence in the estimator structure.

Conclusion

Many physical processes are described by a signal-dependent observation model. It has been shown that, in such cases, ignoring the signal-dependence for purposes of designing estimators of the signal may result in severe penalties in terms of estimation error. Therefore, optimal

estimators which include the signal-dependent structure were derived. Specifically, these were ML estimates, which include no prior knowledge of signal statistics, and MAP estimates, which assume prior knowledge of the signal probability density. The latter estimate was derived for the Gaussian, Rayleigh, and folded Gaussian density functions. The performance of these estimators was then investigated by Monte Carlo simulation. As expected, inclusion of signal-dependence in the estimator structure resulted in improved estimator performance.

Appendix

Bayesian estimators are those estimators which serve to minimize the Bayes risk, where the Bayes risk is the expected cost of estimation based on some cost function. For example, minimum mean-square error is achieved when the cost is proportional to the square of the estimation error, i.e., when the cost function is a parabola. The MAP estimator is a Bayesian estimator based on the uniform cost function shown in Fig. 7.[2] The cost for no error is zero (as it is for some Δ region about no error), and the cost of any other error is uniform (all errors are weighted equally).

It can be shown[2] that, under certain conditions, the optimal Bayes estimate is invariant for a variety of cost functions, and is equal to the minimum mean-square error estimate. These conditions are: (1) the cost function is

convex, (2) the cost function is symmetrical, (3) the a posteriori probability density, $p(s|r)$, is symmetrical, and (4) $\lim_{s\to\infty} C(s)p(s|r) = 0$, where $C(s)$ is the cost function with argument s. Condition (4) is simply a requirement that the a posteriori density goes to zero faster than the cost function increases. Viterbi[14] has shown that the uniform cost function satisfies these conditions. When the prior signal density, $p(s)$, is assumed Gaussian, then clearly $p(s|r)$ is symmetrical, as required in condition (3). Thus, for this case we have the optimal Bayesian estimate, and it is the estimate which yields the minimum mean-square estimation error.

References

[1]J. B. Thomas, An Introduction to Statistical Communication Theory (Wiley, New York, 1969).

[2]H. L. Van Trees, Detection, Estimation, and Modulation Theory, Part 1 (Wiley, New York, 1968).

[3]H. C. Andrews and B. R. Hunt, Digital Image Restoration (Prentice-Hall, Englewood Cliffs, New Jersey, 1977).

[4]J. F. Walkup and R. C. Choens, "Image Processing in Signal-Dependent Noise," Optical Engineering 13, 258-266 (1974).

[5]J. C. Mallison, "Tutorial Review of Magnetic Recording," Proc. IEEE 64, 196-223 (1976).

[6]J. W. Goodman and J. F. Belsher, "Fundamental Limitations in Linear Invariant Restoration of Atmospherically Degraded Images," Proc. SPIE 75, 141-154 (1976).

[7]F. Naderi, "Estimation and Detection of Images Degraded by Film-Grain Noise," Ph.D. thesis (University of Southern California, September, 1976), USC Image Processing Institute report 690.

[8]B. R. Hunt, "Bayesian Methods in Nonlinear Digital Image Restoration," IEEE Trans. on Computers C-26, 219-229 (1977).

[9]H. J. Zwieg, E. B. Barrett and P. C. Hu, "Noise-Cheating Image Enhancement," J. Opt. Soc. Am. 65, 1347-1353 (1975).

[10]A. A. Sawchuk, private communication.

[11]T. G. Stockham, Jr., "Image Processing in the Context of a Visual Model," Proc. IEEE 60, 828-842 (1972).

[12]A. M. Mood, F. A. Graybill and D. C. Boes, Introduction to the Theory of Statistics (McGraw-Hill, New York, 1974).

[13]R. J. Arguello, "Encoding, Transmission and Decoding of Sampled Images," Symposium on Sampled Images, 1971, Perkin-Elmer Corp.

[14]A. J. Viterbi, Principles of Coherent Communication (McGraw-Hill, New York, 1966).

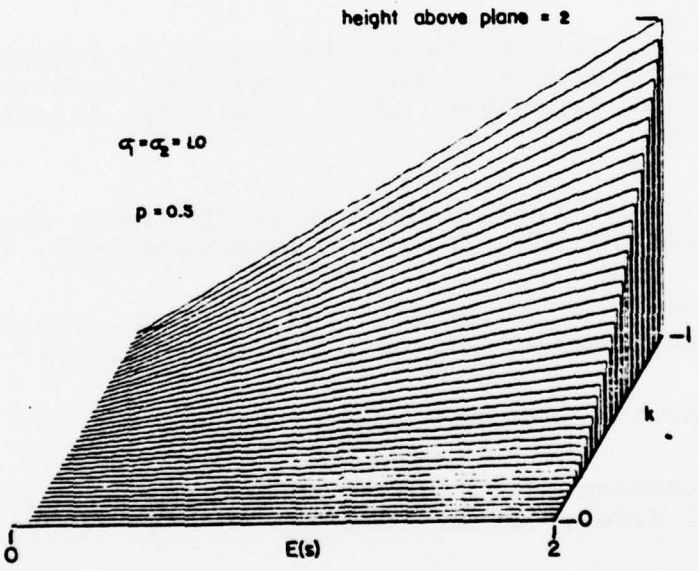


Figure 1. Mean square estimation error lower bound for the mismatched case, p = 1/2.

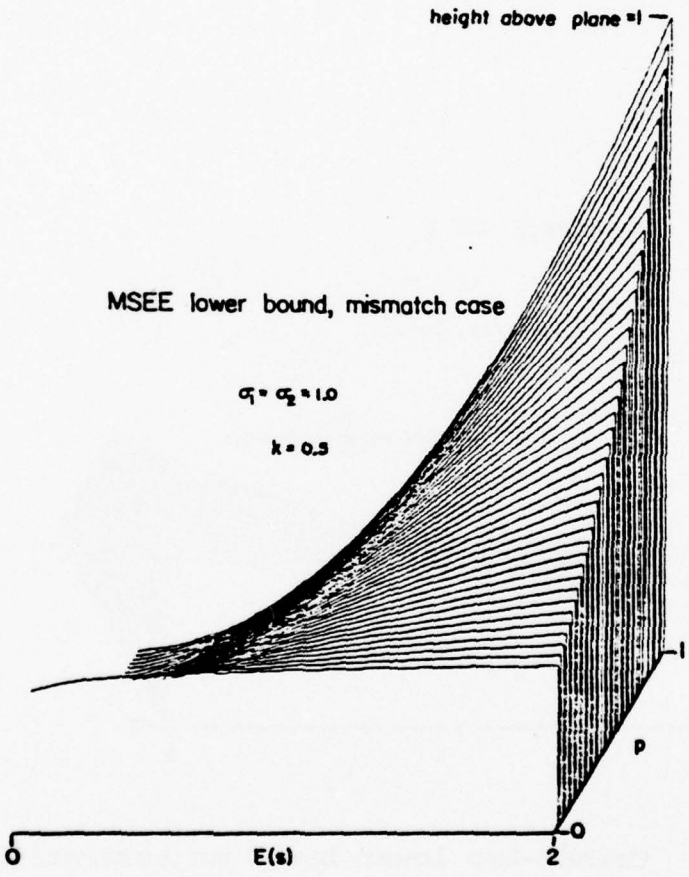


Figure 2. Mean square estimation error lower bound for the mismatched case, k = 1/2.

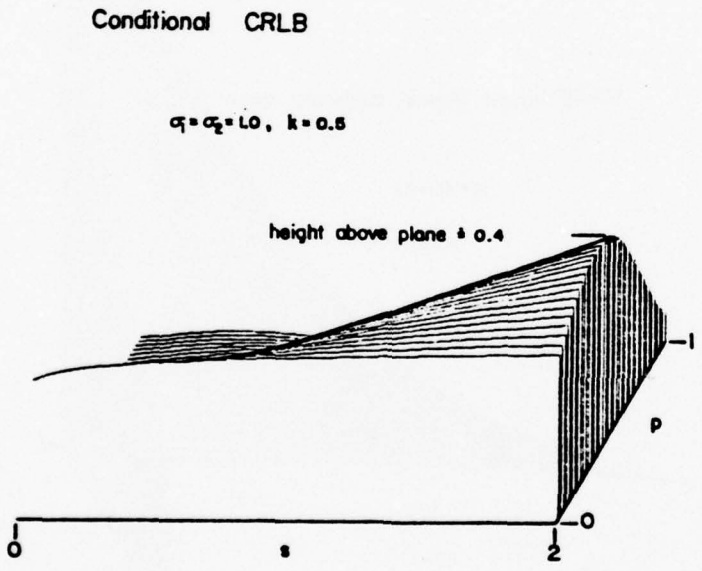


Figure 3. Cramér-Rao lower bound on estimation error for the signal-dependent measurement model, k = 1/2.

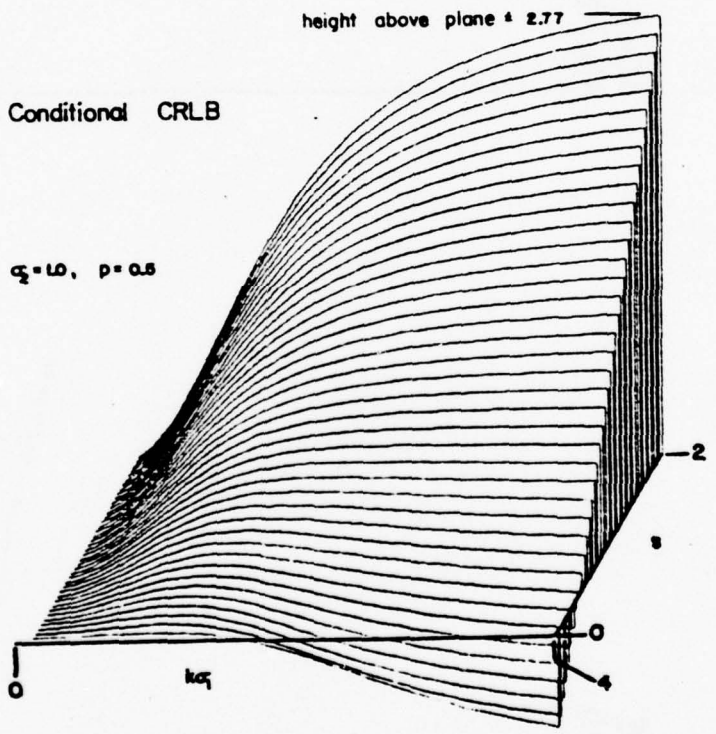


Figure 4. Cramér-Rao lower bound on estimation error for the signal-dependent measurement model, p = 1/2.

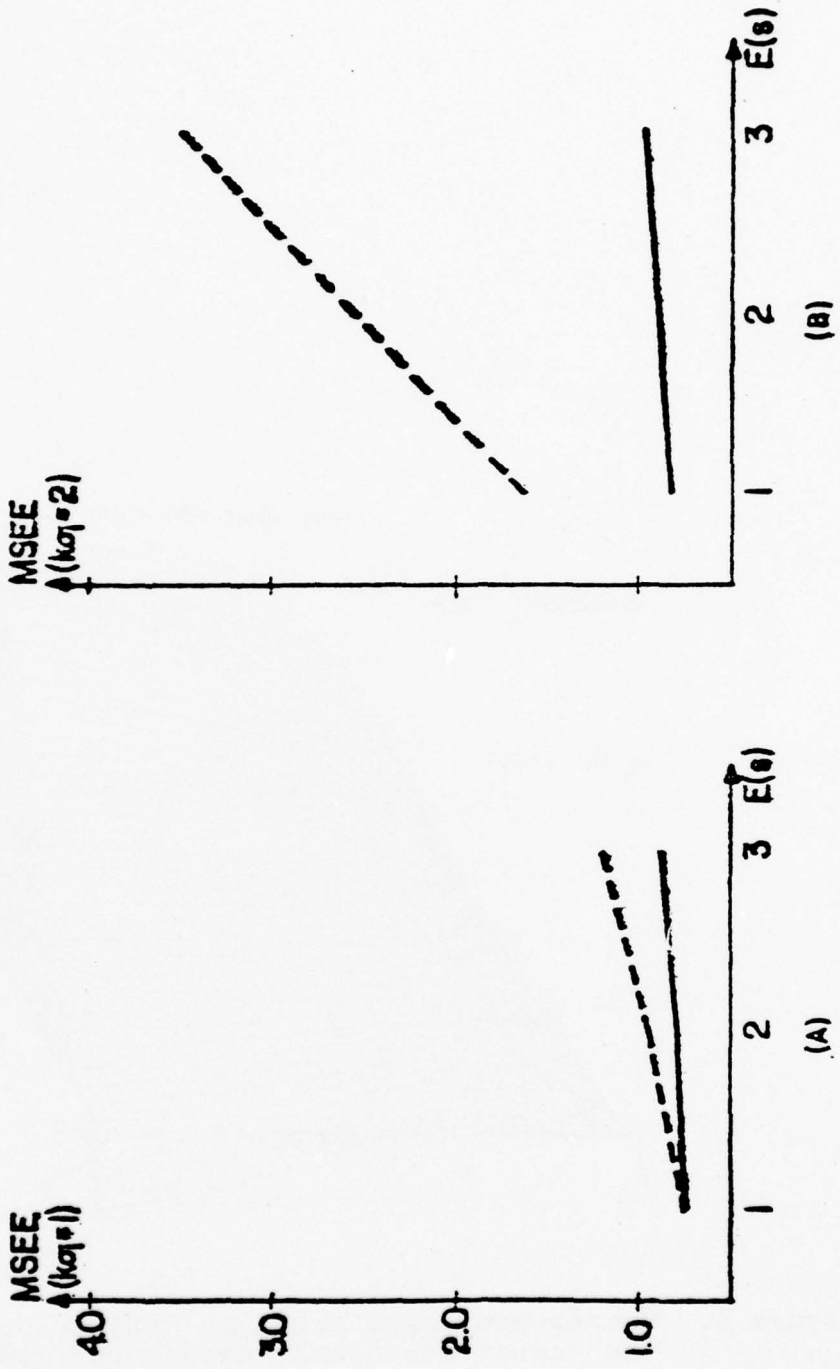


Figure 5. Mean square estimation error for the MAP estimator, as a function of the signal mean E(s), with a) $k\sigma_1 = 1$ and b) $k\sigma_1 = 2$. The solid line is the signal-dependent estimator error and the dashed line is the mismatched estimator error.

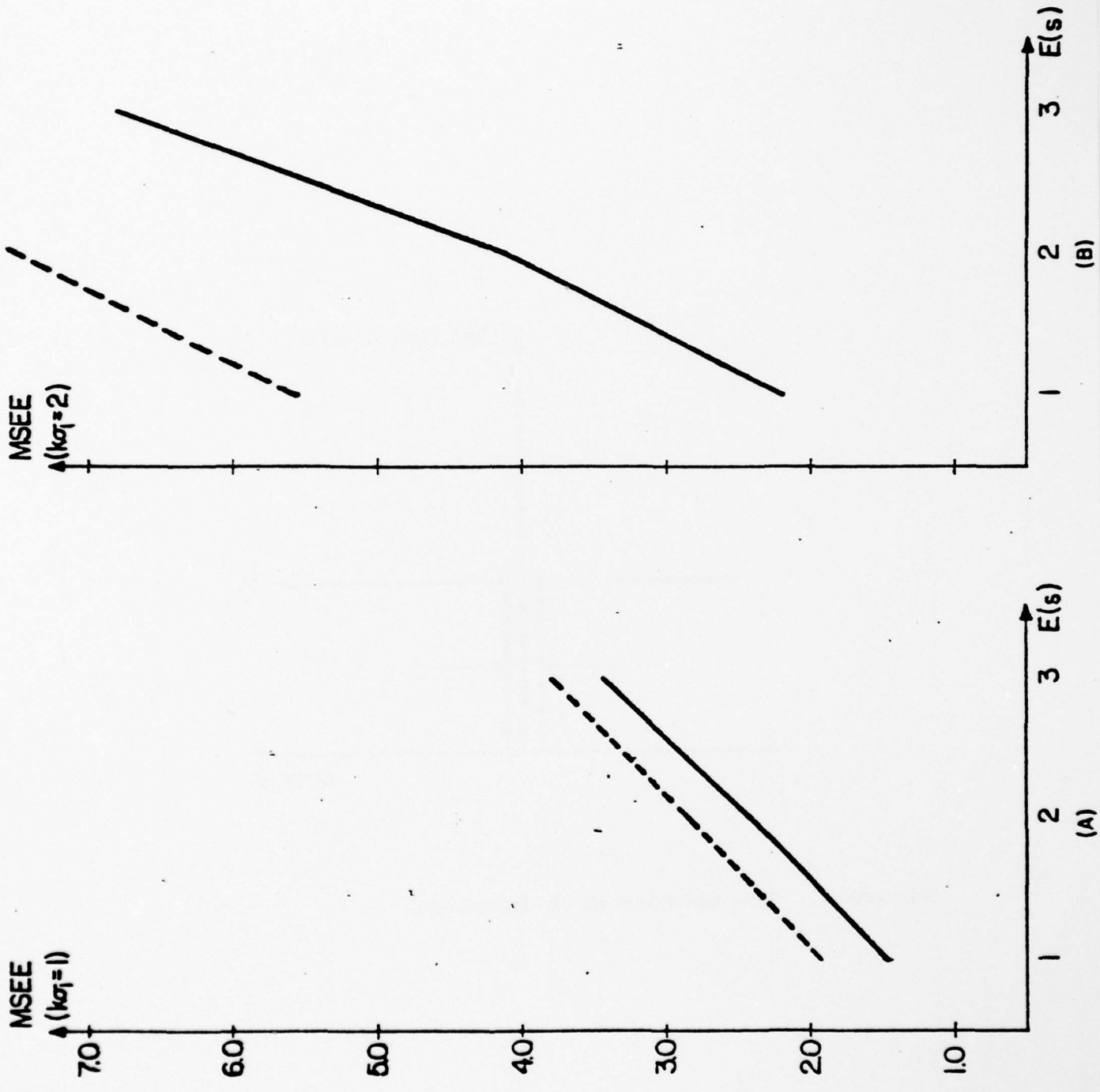


Figure 6. Mean square estimation error for the ML estimator, as a function of the signal mean E(s), with a) $k\sigma_1=1$ and b) $k\sigma_1=2$. The solid line is the signal-dependent estimator error and the dashed line is the mismatched estimator error.

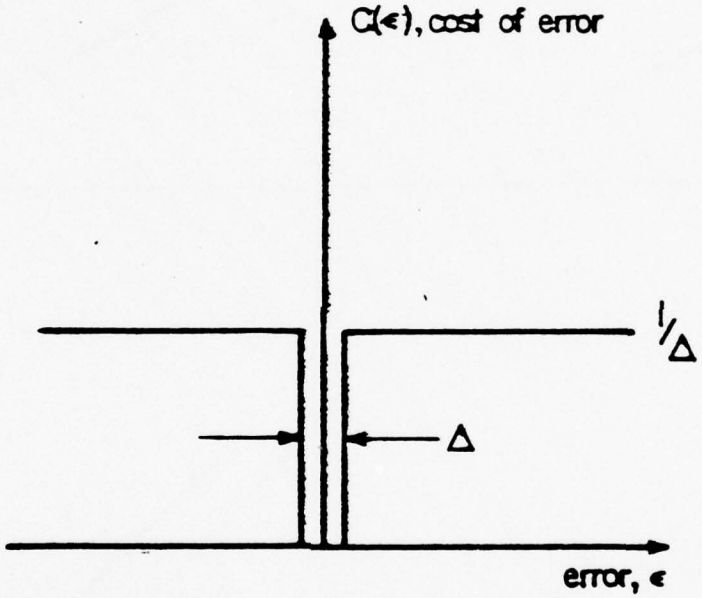


Figure 7. The uniform cost function.

RESEARCH

on

PATTERN RECOGNITION

T. Newman

DEPARTMENT OF MATHEMATICS

TEXAS TECH UNIVERSITY

Abstract

A survey of ongoing research into the existence invariants and relative invariants for application in pattern recognition is presented. A mathematical formalism is developed and a complete characterization of the relative invariants is given.

Introduction

The importance of group theory as a tool to be exploited in modelling a variety of perceptual phenomena has been demonstrated by a number of writers [2,11,12,18]. Although the influence of group theory is implicit in much of the literature on pattern recognition [1,8,13,16,20], relatively few instances can be found in which explicit utilization of group theory is the central theme [4,6,7,19]. Without exception, group theory has been used to effectively model some aspect or feature which is invariant under transformation and to exploit this invariance in performing the recognition function. However, no definitive study has been made of transformational invariance and no general model has been introduced which attempts to formalize the concept of invariance as it relates to pattern recognition. This is indeed strange in view of the relatively advanced state of the theory of invariants within group theory [10,15,21].

In the following we formulate a general model in which many problems in pattern recognition may be cast in a natural fashion. We discuss representations of patterns as functions defined on a group and proceed to investigate the existence of invariant functionals.

The Model

Let Ω denote a set of objects called patterns and assume that G is a group of transformations which act on Ω on the left. For $w \in \Omega$ and $g \in G$ we denote by gw the image of w under the transformation g. Also, for $g_1, g_2 \in G$ we denote their product or composition by $g_1 g_2$. Action on the left is then given by the identity

$$(g_1 g_2)w = g_1(g_2 w), \tag{1}$$

for $g_1, g_2 \in G$ and $w \in \Omega$.

We now assume that our ability to "recognize" and/or otherwise "classify" patterns is obtained via measurements performed upon individual patterns. Such measurements can take values of a quite general nature, although the usual situation will result in a vector of real numbers. Accordingly, we define a measurement function to be a mapping $R: \Omega \to V$, where V is a suitable set of permissible values. We shall later assume that V is a real finite-dimensional vector space. We say that a measurement function $R: \Omega \to V$ is invariant provided that $R(gw) = R(w)$ for all $w \in \Omega$ and $g \in G$. Observe that an invariant measurement does not distinguish between the various members of an orbit $[w] = \{gw | g \in G\}$, being constant on each such orbit.

More generally, we say that a measurement $R: \Omega \to V$ is relatively invariant provided that $R(gw) = \rho(g)R(w)$. Here ρ is a homomorphism of G into a group of transformations on V and is called the modulus of R. As a matter of practice, we are interested in the case in which V is a finite dimensional vector space and ρ

is a representation of G in the group GL(v) of invertible linear transformations on V. Note that a relative invariant not only depends upon the orbit of w but is also sensitive to "position" within the orbit.

In applications one must solve simultaneous equations $R_\alpha(w) = R_\alpha^m$ involving a number of invariants $\{R_\alpha\}$ and associated actual measurements $\{R_\alpha^m\}$ to classify the orbit of w and then solve similar equations $\rho_\beta(g)R_\beta(w) = R_\beta^m$ involving relative invariants to determine position within the orbit. Hence we see that the question of existence of invariants and relative invariants become of paramount importance.

Representations

In order to pursue the question of existence of invariants we find the need of considerably more structure than we have assumed at this point. It is somewhat surprising that this additional structure can be imposed on the transformation group and need not involve restrictive assumptions about the space of patterns. Since the transformation groups that are typically encountered are quite rich in structure, we find ourselves in an advantageous situation.

Let us briefly digress. Suppose that X is any set and that the group G acts on X on the left. Let $f: X \to Y$ be a mapping of X to some set Y. Then for any g G we may define a new mapping $gf: X \to Y$ given by

$$(gf)(x) = f(g^{-1}x), \quad x \in X. \tag{2}$$

Note that appearance of g^{-1}, rather than g, is a convenience which makes certain formulae more natural for later use. We easily verify that

$$g_1(g_2 f) = (g_1 g_2) f \tag{3}$$

so that if F is a set of functions such that $gf \in F$ for all $f \in F$, then (2) defines an action of G on the left of F.

Now let $R: \Omega \to V$ be a given measurement function. For each $w \in \Omega$ we may define a function $w^r: \Omega \to V$ as follows:

$$w^r(x) = R(x^{-1}w), \quad x \in G \tag{4}$$

The correspondence $r: w \to w^r$ thus defines a mapping of Ω into the set F(G,V) of functions from G to V. Now for fixed $g \in G$ we see that for all $x \in G$, $(gw^r)(x) = w^r(g^{-1}x) = R((g^{-1}x)^{-1}w) = R((x^{-1}g)w) = R(x^{-1}(gw)) = [(gw)^r](x)$. That is,

$$gw^r = (gw)^r, \tag{5}$$

for all $w \in \Omega$ and $g \in G$. Equation (5) establishes the desired connection between our patterns and the V-valued functions on G. We define a representation of Ω on V to be a map $r: \Omega \to F(G,V)$ which satisfies (5). Such a representation allows a <u>concrete</u> interpretation of patterns as suitable functions defined on the group.

We have the following

<u>Theorem 1</u>. The representations r of Ω on V correspond one-to-one to the measurement functions $R: \Omega \to V$. The correspondence is given via $r \leftrightarrow R$ if and only if

$$w^r(x) = R(x^{-1}w), \; x \in G, \; w \in \Omega.$$

Proof: We have already seen that each measurement R defines a representation. Now, let $w \to \tilde{w}$ be a representation of Ω on V and let us set $R(w) = \tilde{w}(1_G)$, where 1_G is the identity element of We must show that w^r as defined by (4) satisfies $w^r = \tilde{w}$. But for $x \in G$ we have $\tilde{w}(x) = (x^{-1}\tilde{w})(1_G) = \widetilde{(x^{-1}w)}(1_G) = R(x^{-1}w) = w^r(x)$, from which $\tilde{w} = \tilde{w^r}$, as desired.

Let us point out that an invariant measurement R is characterized by the condition that each w^r is a constant function, which on the surface seems somewhat uninteresting. This is a deceptive simplification, however, as will be apparent later. Similarly, if R is a relative invariant with modulus ρ, we see that $w^r(x) = \rho(x^{-1})R(w)$.

Before proceeding to pursue the existence of invariants, it seems appropriate to further accent the importance of relative invariants by demonstrating one of their fundamental properties. Let $R: \Omega \to V$ be a relative invariant with modulus ρ. Suppose that $w_1, w_2 \in \Omega$ and that $R(w_1) = R(w_2)$. Then for any $g \in G$ we have $R(gw_1) = \rho(g)R(w_1) = \rho(g)R(w_2) = R(gw_2)$. Thus,

$$R(w_1) = R(w_2) \text{ implies } R(gw_1) = R(gw_2).$$

It is somewhat interesting to note that the condition above is a complete characterization of relative invariants as is shown in the following:

Theorem 2. In order that $R: \Omega \to V$ be relatively invariant it is necessary and sufficient that

$$R(w_1) = R(w_2) \text{ implies } R(gw_1) = R(gw_2)$$

for all $g \in G$.

Proof: Necessity has already been shown. Conversely, suppose that $R: \Omega \to V$ satisfies the stated condition. We must construct a homomorphism of G into the group Sym(V) of transformations on V. For $v = R(w) \in v$ and $g \in G$, let us define $\rho(g)v = R(gw)$. We note that this definition does not depend on w for if also $v = R(w')$ then $R(gw') = R(gw)$, by the property of R. If $v \in V$ is <u>not</u> of the form $v = R(w)$ then set $\rho(g)v = v$. We easily verify that each $\rho(g) \in \mathrm{Sym}(V)$. Also,

$$\rho(g_1)\rho(g_2)v = \rho(g_1)\rho(g_2)R(gw) = \rho(g_1)R(g_2w) = R(g_1g_2w) = \rho(g_1g_2)v$$

in case $v = R(w)$ and $\rho(g_1)\rho(g_2)v = v = \rho(g_1g_2)v$ otherwise.

Thus, $\rho(g_1)\rho(g_2) = \rho(g_1g_2)$ so that ρ is indeed a homomorphism.

Finally, by definition of $\rho(g)$ we see that $R(gw) = \rho(g)R(w)$ for all $g \in G$ and $w \in \Omega$ so that R is a relative invariant with ρ as modulus.

Invariants and Relative Invariants

As previously stated, we may impose additional structure by invoking restrictions on the transformation group. In the following we assume that V is a real vector space of finite dimension and that G is a locally compact topological group. Such a group admits a left invariant integral, called left Haar measure [15], and the integration theory for such groups is well established.

The fundamental technique for construction of invariants will be the computation of average values over the entire group G.

This technique was exploited by Pitts and McCulloch [18] in their classic work on the perception of audio and visual forms. It appears also in the classical theory of group representation [21] and is prevalent in modern analysis [10,15]. Group averaging has been used as a tool in pattern recognition in a relatively few instances, implicitly in [1,13] and explicitly in [5,6,7].

Now, let y denote the left Haar measure on G and let $f: G \to V$. We define the mean value of f, provided it exists, by

$$M(f) = \lim_{K \uparrow G} \frac{1}{M(K)} \int_K f \, dM, \tag{6}$$

where K is a compact subset of G and the limit is taken as K increases. Note that K compact implies that $g^{-1}K$ is compact and that $y(K) = y(g^{-1}K)$. This together with the fact that $\int_K gf dy = \int_{g^{-1}K} f \, dy$, $g \in G$, shows the following:

Lemma 1. If $M(f)$ exists then for any $g \in G$, $M(gf)$ exists and $M(f) = M(gf)$.

We denote by $L(V)$, or simply L, the set of all $f: G \to V$ for which $M(f)$ exists. We have the following:

Lemma 2. $L(V)$ is a linear space on which G acts as the left as a group of linear transformations. Moreover, M is an invariant linear transformation of $L(V)$ into V.

More generally, let ρ be a representation of G in the group $GL(V)$ of invertible linear transformations on V. We form the weighted average of $f: G \to V$, provided the limit exist, as follows:

$$M_\rho(f) = \lim_{K \uparrow G} \frac{1}{\mu(K)} \int_K \rho(x) f(x) d\mu(x), \tag{7}$$

where K is compact, as above. The set of functions for which

$M_\rho(f)$ exists will be denoted by $L(V,\rho)$, or simply $L(\rho)$. Note that by the substitution $y = g^{-1}x$ we obtain

$$\int_K \rho(x) gf(x)\, dy(x) = \int_K \rho(x) f(g^{-1}x)\, dx(x) = \int_{g^{-1}K} \rho(gy) f(y)\, dy(y) =$$

$$\rho(g) \int_{g^{-1}K} \rho(y) f(y)\, dy(y).$$ It follows immediately that

$$M_\rho(gf) = \rho(g) M_\rho(f), \tag{8}$$

for all $g \in G$, $f \in L(V,\rho)$. We conclude:

Theorem 3. $L(V,\rho)$ is a linear space on which G acts as a group of linear transformations. Also, M_ρ is a linear mapping of $L(V,\rho)$ into V which is relatively invariant with modulus ρ.

Let us define $\rho' : G \to GL(V)$ by $\rho'(x) = \rho(x^{-1}) = (\rho(x))^{-1}$. We have $\rho'(xy) = \rho'(y)\rho'(x)$, so that ρ' is a dual homomorphism. For $f \in L(V)$ we may consider the product $\rho' f$ given by $(\rho' f)(x) = \rho'(x) f(x)$, $x \in G$. Since $\rho(x)\rho'(x) = 1_V$ we see that $\rho' f \in L(V,\rho)$ whenever $f \in L(V)$. Similarly, $f \in L(V)$ implies $\rho f \in L(V,\rho)$.

We evidently then can use ρ and ρ' as multipliers to pass back and forth between $L(V)$ and $L(V,\rho)$. Thus:

Lemma 3: The map $f \to \rho' f$ is a linear isomorphism from $L(V)$ onto $L(V,\rho)$. Moreover, $M(f) = M_\rho(\rho' f)$.

Although this shows that the linear structure of $L(V)$ and $L(V,\rho)$ are no different, it is important to observe that they are quite different with respect to the action of G.

We may now state sufficient conditions for the existence of invariants and/or relative invariants for the pattern space Ω. Quite simply, if $R: \Omega \to V$ is such that each $w^r \in L(V,\rho)$ then

we obtain a relative invariant $\bar{R}$ by defining

$$\bar{R}(w) = M_\rho(w^r) \tag{9}$$

We see that $\bar{R}(gw) = M_\rho((gw)^r) = M_\rho(g\ w^r) = \rho(g)\ M_\rho(w^r) = \rho(g)\bar{R}(w)$, as desired. We obtain the corresponding result for invariants in the special case in which ρ is the trivial representation, $\rho(g) \equiv 1_V$.

Recall that if $R: \Omega \to V$ is relatively invariant with modulus ρ, then we may write $w^r(x) = \rho(x^{-1})R(w)$ for all $w \in \Omega$, $x \in G$. Thus, we have for each compact subset K of G, $\int_K \rho(x) w^r(x)\, d\mu(x) = \int_K R(w)\, dy(x) = \mu(K)\ R(w)$. Comparison with (7) shows that $M_\rho(w^r)$ exists and is equal to $R(w)$. This shows that each $w^r \in L(V,\rho)$ as well as the identity $M_\rho(w^r) = R(w)$. We have thus shown:

Theorem 4. If $R: \Omega \to V$ is such that each $w^r \in L(V,\rho)$, then $\bar{R}(w) = M_\rho(w^r)$ defines a relative invariant $\bar{R}$ with modulus ρ. Conversely, every relative invariant is precisely of this form, since if R is relatively invariant with modulus ρ, then each $w^r \in L(V,\rho)$ and $\bar{R} = R$.

The above may be paraphrased by saying that the construction of relative invariants with a given modulus ρ is equivalent to the construction of a representation $w \to w^r$ of Ω on V such that each $w^r \in L(V,\rho)$. Observe that, in particular, we have shown in a strict sense that every relative invariant is a weighted average over the entire group G.

The result in Theorem 4 gives valuable insight to the nature of invariants and relative invariants. Nevertheless, it is less than satisfying in certain ways. In the first place, it gives no clue as to how to construct a suitable R, although it can certainly eliminate a number of choices, Consequently, it is not a true existence theorem in the sense that for a given application it does not actually produce an invariant. Moreover, there are many examples of invariants which occur in natural ways but are not presented in the form given above (although they are necessarily equivalent to such a form).

Existence of Invariants

The consideration of the group average in the proceeding section led to the existence of relative invariants and is applicable in any situation where each w^r belongs to a class of functions for which such an average exists. This is the case, for example, when the class of functions is almost periodic (in the sense of J. von Neumann [10]. It is, however, applicable in a wider variety of cases, namely those in which set B(G) of bounded real valued functions admits an invariant mean, in the following sense:

Definition: An invariant mean M on the class B(G) of bounded real valued functions on G is a real linear functional M on B(G) which is invariant under the action of G on B(G) and satisfies

$$\inf f \leq M(f) \leq \sup f \ , \ f \in B(G). \tag{10}$$

Let V^* denote the dual space of V and for each $f \in B(G,V)$, the bounded functions from G to V, and for each $v^* \in V^*$ we observe that $v^* \circ f \in B(G)$.

Lemma 4. Let $M_0 \in (B(G))^*$, the dual of B(G). There exists a unique $M \in \mathrm{Lin}(B(G,V),V)$ such that

$$v^* \circ M = M_0 \circ v^* \tag{11}$$

for all $v^* \in V^*$.

Proof: It is clear that any M satisfying (11) is unique. Let $v_1, v_2, \ldots, v_n$ be a basis in V and $v_1^*, v_2^*, \ldots, v_n^*$ a dual basis in V^*, so that $\langle v_i^*, v_j \rangle = \delta_{ij}$. Let us define $M: B(G,V) \to V$ by

$$M(f) = \sum_{i=1}^{n} M_0(v_i^* \circ f) v_i \ , \quad f \in B(G,V). \tag{12}$$

Then for any $v^* \in V^*$ we have $v^* \circ M(f) = \langle v^*, M(f) \rangle = \sum_{i=1}^{n} \langle v^*, v_i \rangle M_0(v_i^* \circ f) = M_0(\sum_{i=1}^{n} \langle v^*, v_i \rangle v_i^* \circ f) = M_0(v^* \circ f)$. That $v^* \circ M = M_0 \circ V^*$, as desired.

Let us observe that for any $A \in \mathrm{Lin}(V,V)$ and $f \in B(G,V)$ we may define the composite Af so that B(G,V) may be considered as a (left) module over Lin(V,V). With this in mind, we observe:

Lemma 5: The linear map $M: B(G,V) \to V$ defined by (11) above is a morphism of B(G,V) to V considered as modules over Lin(V,V).

Proof: For any $A \in \mathrm{Lin}(V,V)$, $v^* \in V^*$ and $f \in B(G,V)$, we have $v^* \circ M(Af) = M_0(v^* \circ Af) = M_0((A^* v^*) \circ f) = A^* v^* \circ M(f) = v^* \circ AM(f)$. Hence, $M(Af) = AM(f)$, completing the proof.

Lemma 6: If $M_0 \in (B(G))^*$ is invariant under G then so is the map M as defined by (11).

Proof: M_0 invariant means that for any $f_0 \in B(G)$ and $g \in G$, we have $M_0(gf_0) = M_0(f_0)$. Thus, if $v^* \in V^*$, $f \in B(G,V)$ then for any $g \in G$ we have $v^* \circ M(gf) = M_0(v^* \circ f) = M_0(g(v^* \circ f)) = M_0(v^* \circ f) = v^* \circ M(f)$. Then, $M(gf) = M(f)$, as desired.

We have thus shown how to "lift" invariant linear functionals on B(G) to invariant linear maps from B(G,V) to V.

Corollary: If G admits an invariant mean M_0 and $R: \Omega \to V$ is such that each $w^r \in B(G,V)$, then

$$\bar{R}(w) = M(w^r) , \tag{12}$$

where M is given by (11), is an invariant measurement.

Proof: $\bar{R}(gw) = M((gw)^r) = M(gw^r) = M(w^r) = \bar{R}(w)$.

We may obtain relative invariants in a similar fashion. However, to remain within the bounded functions, we restrict our attention to unitary representations of G.

Let us suppose that M_0 is a given invariant mean on the class of bounded function on G and that M is the lifted map defined by (11) above. Also, let ρ be a given unitary representation of G in GL(V). Observe then that for each $f \in B(G,V)$ we have also $\rho f \in B(G,V)$, where $(\rho f)(g) = \rho(g)f(g)$, $g \in G$. Now, let $R: \Omega \to V$ be a given measurement function such that each $w^r \in B(G,V)$. Since this simply means that the values of R on the orbit of w are bounded, this is not deemed to be a serious restriction.

With this in mind, let us note that $\rho(gw)^r = \rho(g)\ g(\rho w^r)$ for all $g \in G$, $w \in \Omega$. To see this, we have, at any $x \in G$, $[\rho(gw)^r](x) = \rho(x)(gw^r)(x) = \rho(x)\ w^r(g^{-1}x) = \rho(g)\ \rho(g^{-1}x)w^r(g^{-1}x) = \rho(g)[g(\rho w^r)](x)$. Also, let us observe that, for fixed g, $\rho(g) \in \mathrm{Lin}(V,V)$ and that M is a morphism of $\mathrm{Lin}(V,V)$-modules. We now define $\bar{R}\colon \Omega \to V$ by the formula

$$\bar{R}(w) = M(\rho w^r)\ ,\ w \in \Omega. \tag{13}$$

Recalling the invariance of M, and the facts above, we see that for $g \in G$, we have $\bar{R}(gw) = M(\rho(gw)^r) = M(\rho(g)g(\rho w^r))$ $= \rho(g)\ M(g(\rho w^r)) = \rho(g)M(\rho w^r) = \rho(g)\bar{R}(w)$.

That is, $\bar{R}$ is a relative invariant and has the given representation ρ as its modulus. We have therefore proved the following remarkable result:

Theorem 5. If $B(G)$ admits an invariant mean M_0, ρ is any unitary representation of G in $GL(V)$, and a non-trivial bounded measurement function $R\colon \Omega \to V$ exists, then there exists a non-trivial relative invariant $\bar{R}\colon \Omega \to V$ with modulus ρ. $\bar{R}$ is given explicitly by

$$\bar{R}(w) = M(\rho w^r)\ , \tag{13}$$

where M is the lift of M_0 to $B(G,V)$.

Note: The appearance of the words non-trivial in the above requires slight explanation. We can clearly define $\bar{M}\colon B(G,V) \to V$ by $\bar{M}(f) = M(\rho f)$ and deduce that $\bar{M}(gf) = \rho(g)\bar{M}(f)$. The fact that $M \not\equiv 0$ gives $\bar{M} \not\equiv 0$. Since $\bar{R}(w) = \bar{M}(w^r)$, we see the sense in which

$\bar{R}$ is non-trivial, i.e., it is the restriction of $\bar{M}$ to the functions $\Omega^r = \{w^r | w \in \Omega\}$. Nevertheless, it could happen that each $\rho w^r \in \ker M$ so that $\bar{R} \equiv 0$ even though $R \neq 0$. This is unlikely and can be ignored if we have some $\rho w^r > 0$. For such $w \in \Omega$ we see that $\bar{R}(w) > 0$.

Summary

We have shown that every set of patterns subject to a transformation group is representable as functions defined on the group and that such representations are implicit in the measurement process. It has also been shown that every relative invariant is equivalent to a weighted average over the group of a measurement on the patterns. Moreover, the existence of suitable many relative invariants have been demonstrated in any situation in which measurements are bounded and the group admits an invariant mean.

References

[1] F. Alt, Digital Recognition by Moments; in Optical Character Recognition, Washington, D.C., Spartan, 1962, pp. 152-179.

[2] E. Cassirer, The Concept of Group and the Theory of Perception, Philosophy and Phenomenological Research, Vol. V, 1944, pp. 1-35.

[3] P. M. Cohn, Lie Groups, Cambridge University Press, 1957.

[4] H. Dirilten, Ph.D. Dissertation, Texas Tech University, 1974.

[5] H. Dirilten and T. G. Newman, Pattern Matching Under Affine Transformations, IEEE Trans. Comp., Vol. C-24, pp. 314-317, 1977.

[6] J. C. Dunn, Continuous Group Averaging and Pattern Classification Problems, SIAM J. Comput., Vol. 2, 1973, pp. 253-272.

[7] J. C. Dunn, Group Averaged Linear Transforms that Detect Corners and Edges, IEEE Trans. Comp., Vol. C-24, 1975, pp. 1191-1201.

[8] R. Duda and P. Hart, Pattern Classification and Scene Analysis, New York, John Wiley and Sons, 1973.

[9] M. O. Hagler, T. G. Newman and H. Dirilten, A Programmable Optical-Digital Scanner-Processor for Automated Two-dimensional Data Analysis, IEEE Trans. Comp. Vol. C-24, pp. 1036-1038, 1975.

[10] E. Hewitt and K. Ross, Abstract Harmonic Analysis I, New York, Academic Press Inc., 1963.

[11] W. C. Hoffman, The Lie Algebra of Visual Perception, J. Math. Psych., Vol. 3, 1966, pp. 65-98.

[12] W. C. Hoffman, The Neutron as a Lie Group Germ and a Lie Product, Quart. Appl. Math., Vol. 25, 1968, pp. 423-440.

[13] M. K. Hu, Visual Recognition by Moment Invariants, IRE Trans. Inform. Thy., Vol. IT-8, 1952, pp. 179-187.

[14] R. B. McGhee, Automatic Recognition of Complex Three-Dimensional Objects from Optical Images, Report AFOSR-TR-74-0090 under contract AFOSR-71-2048, National Technical Information Service, Oct. 1973.

[15] L. Nachbin, The Haar Integral, Princeton, N.J., Van Nostrand Inc., 1965.

[16] G. Nagy, State of the Art in Pattern Recognition, Proc. IEEE, Vol. 26, 1968.

[17] T. G. Newman and H. Dirilten, A Nonlinear Transformation

[18] W. Pitts and W. S. McCulloch, How we know Universals - The Perception of Auditory and Visual Forms, Bull. Math. Biophysics, Vol. 9, 1967, pp. 127-147.

[19] J. M. Richardson, Pattern Recognition and Group Theory, in Frontiers of Pattern Recognition, New York, Academic Press, 1972, pp. 453-477.

[20] A. D. Van der Lugt, Signal Detection by Complex Spatial Filtering, IEEE Trans. Info. Thy., Vol. IT-10, 1964.

[21] H. Weyl, The Classical Groups, Princeton University Press, Princeton, N.J., 1946.

[22] E. Wong and J. A. Steppe, Invariant Recognition of Geometric Shapes, in Methodologies in Pattern Recognition, New York.

REVIEW OF RESEARCH

in

ELECTRONICS AND RELATED AREAS

RESEARCH FACULTY

in

ELECTRONICS AND RELATED AREAS

SYSTEMS

Name	Title	Area	Office	Phone
K.S. Chao	Assoc. Prof. EE	Nonlinear Circuits and Systems	150A-EE	742-3469
R.A. DeCarlo	Lect.-EE	Stability Theory	258-EE	742-3528
R.M. DeSantis	Visiting Assoc. Prof.-EE	Math. Syst. Theory	258-EE	742-3528
D.L. Gustafson	Assoc. Prof.-EE	Microprocessors	150B-EE	742-3530
M.O. Hagler	Prof.-EE	Optical Signal Proc.	103B-EE	742-3470
T.F. Krile	Visiting Assoc. Prof.-EE	Optical Systems	151-EE	742-3500
S.R. Liberty	Assoc. Prof. EE & Stat.	Stochastic Control and Estimation	201B-EE	742-3441
J. Murray	Research Assoc.-EE	Multi-dimensional Digital Filters	258-EE	742-3528
T. Newman	Assoc. Prof. Math & C.S.	Pattern Recognition	1107-BA	742-2571
C.T. Pan	Lect.-EE	Numerical Analysis & Nonlinear Cir.	258-EE	742-3528
J. Prabhakar	Assoc. Prof. EE	Communications	104-EE	742-3506
R. Saeks	Prof.-EE and Math	Circuits and Systems	258A-EE	742-3528
L. Tung	Lect.-EE	Math. Systems Theory	258A-EE	742-3528
J. Walkup	Assoc. Prof. EE	Optical Systems & Communications	260B-EE	742-3500

PHYSICAL ELECTRONICS

M. Gundersen	Assoc. Prof. EE	Quantum Electronics	260A-EE	742-3501
A. Kwatra	Lect.-EE	Optics & Quantum Electronics	205-EE	742-3502

W. Portnoy	Prof.-EE	Solid State & Bio-Med. Elec.	152-EE	742-3532
J. Reichert	Assoc. Prof. EE	Optics & Quantum Electronics	203-EE	742-3502
F. Williams	Asst. Prof. EE.	Interaction of Light with Matter	258B-EE	742-3501

ELECTROMAGNETICS

R. Cross	Vis. Prof. EE	Plasma	103C-EE	742-3468
M. Hagler	Prof.-EE	Plasma	103B-EE	742-3470
M. Kristiansen	Horn Prof. EE	Plasma	103A-EE	742-3468
E. Kunhardt	Asst. Prof. EE	Nonlinear Phenomena	260C-EE	742-3545
T. Trost	Assoc. Prof. EE	Antennas & Propagation	102-EE	742-3505

POWER

T. Burkes	Assoc. Prof. EE	Power Conditioning	105C-EE	742-3533
J. Craig	Prof.-EE	Electro-Mech. Devices	101-EE	742-3529
M. Kristiansen	Horn Prof. EE	High Power Switching	103A-EE	742-3468
E. Kunhardt	Asst. Prof. EE	High Power Switching	260C-EE	742-3545
S. Liberty	Assoc. Prof. EE	Solar Energy	201B-EE	742-3441
J. Reichert	Assoc. Prof. EE	Solar Energy	203-EE	742-3502

ACTIVE GRANTS AND CONTRACTS

in

ELECTRONICS AND RELATED AREAS

Systems

Principal Invest.	Agency	Title	Duration	Annual Funding
Portnoy	NASA	Miniaturized Bio-Med Telemetry ...	1 yr.	29,940
Saeks	AFOSR	Resolution Space...	1 yr.	23,129
Liberty	ONR	Statistical Performance Analysis...	1 yr.	30,000
Saeks/Gustafson	ONR	Fault Analysis...	1 yr.	30,000
Saeks/Chao	NSF	Semi-Analytic Methods...	2 yrs.	15,000
Walkup/Hagler	AFOSR	Space-Variant Optical Systems	1 yr.	73,447
Saeks/Levan	AFOSR	Symp. on Oper. Thy. of Networks and Systems	1 yr.	4,468
Asher	AFWL	Estimation in Adaptive Optics	1 yr.	44,330
Asher	RADC	Phased Array Antennea Analysis	1 yr.	35,000
Asher	SORF	Nonlinear Estimation and Detection	1 yr.	5,500
Saeks	ONR	Assoc. Joint Services Electronics Prog.	1½ yrs.	133,333
		Total Annual Funding in Systems		$427,147

Physical Electronics

Gundersen	SORF	Studies in Transient Discharges	1 yr.	8,000
Gundersen	ERDA	Laser Research	1 yr.	75,000
Gundersen	NSF	Inovative Infrared Detector	2 yrs.	17,500
Reichert	AFOSR	Analysis of Unstable Optical Resonators	1 yr.	30,000

Williams/ Gundersen	AFOSR	Studies in Transient Discharges	1 yr.	34,396
Williams	Research Corp.	Driven Raman Proc.	1 yr.	10,000
Kunhardt	NSF	Undergraduate Res. Participation	1 yr.	19,480
Portnoy	SORF	High Temp. Elec.	1 yr.	5,500
		Total Annual Funding in Physical Electronics		$200,416

Electromagnetics

Kristiansen	SORF	Toroidal Plasma Facility	1 yr.	11,000
Kristiansen/ Hagler	NSF	RF Plasma Heating	1 yr.	37,044
Kristiansen	AFOSR	Dense Plasma Heating and Rad. Gen.	1 yr.	99,918
Trost	NSF	Radio Bursts from Severe Storms	2 yrs.	21,250
Trost	AFOSR	Radio Propagation via Transponder	1 yr.	9,950
		Total Annual Funding in Electromagnetics		$179,162

Power

Craig	TPL	Power System Studies	1 yr.	8,000
Burkes	ERDA	Laser Research	1 yr.	25,050
Burkes	ERDA	E Beam Laser Support	1 yr.	15,250
Kristiansen	AFOSR	High Power Switch Dev.	1 yr.	50,000
Craig	AFAPL	Pulse Power Loading	1 yr.	29,000
Kristiansen	ERDA	Surface Flashover Mech.	1 yr.	59,459

Burkes	NSWC	High Power Switches	1 yr.	39,841
Kristiansen	EPRI	RF Heating and Confinement	1 yr.	3,000
Reichert	ERDA	Crosbyton Solar Power Project	1 yr.	500,000*
		Total Annaul Funding in Power		$729,600
		TOTAL ANNUAL FUNDING		$1,536,325

*The Department of Electrical Engineering is the prime contractor on the Crosbyton Solar Power Project which is funded at about $1,500,000 annually. Of this amount about $500,000 is spent in the department with the remainder spent in other departments at Texas Tech and/or subcontracted.

RESEARCH LABORATORIES

in

ELECTRONICS AND RELATED AREAS

SYSTEMS

Computer Laboratories:

CDC 1604 facility: hands-on facility for both education and research.....108-EE

Hybrid Computer facility: minis, micros, and analog facilities.....162-EE

Bio-medical Systems: includes instrumentation and microprocessor application facilities.....215-EE

Circuits and Systems Laboratory: the think tank.....258-EE

Optical Systems Laboratories:

Holographic Optics: primarily used for multiplex holography research.....110-EE

Optical Signal Processing: research in optical and digital image processing.....216-EE

PHYSICAL ELECTRONICS

Laser Laboratory: infrared laser research.....262-EE

Integrated Circuit Laboratory: fabrication facility for SSI and special purpose devices.....209-EE

Laser Laboratory: interaction of light with matter.....260-EE

ELECTROMAGNETICS

Plasma Laboratories:

Laser/Plasma facility: plasma heating via laser plasma interaction.....113-EE

Tokamak facility: radio frequency heating of toroidal plasmas.....117-EE

Electromagnetics Laboratory: nonlinear wave studies.....111-EE

Antenna Laboratory: radio meteorology and ionspheric studies.....West of the Medical School

POWER

High Voltage Laboratory: pulsed power studies.....North of Textile Bldg.

Solar Energy Laboratory: another think tank.....205-EE

High Power Switching Laboratory: electron beam initiated spark gap.....Trailer west of EE Bldg.

PUBLICATION ACTIVITY

in

ELECTRONICS AND RELATED AREAS

Systems - Refereed Publications

Saeks, R., and S.R. Liberty, Rational Fault Analysis, New York, Marcel Dekker, 1977.

Ransom, M.N., and R. Saeks, "A Functional Approach to Fault Analysis of Linear Systems" in Rational Fault Analysis (ed. R. Saeks and S.R. Liberty), New York, Marcel Dekker Inc., 1977, pp. 124-134.

Liberty, S.R., Tung, L., and R. Saeks, "Fault Prediction - Toward a Mathematical Theory" in Rational Fault Analysis (ed. R. Saeks, and S.R. Liberty), New York, Marcel Dekker, 1977, pp. 135-142.

DeCarlo, R.A., Murray, J., and R. Saeks, "Multivariable Nyquist Theory", Int. Jour. on Cont., Vol. 25, pp. 657-675, (1977).

Chao, K.S., and R. Saeks, "Continuation Methods in Circuit Analysis", IEEE Proc., Vol. 65, pp. 1187-1194, (1977).

DeCarlo, R.A., and R. Saeks, "The Encirclement Condition: An Approach Using Algebraic Topology", Int. Jour. on Cont., Vol. 26, pp. 279-287, (1977).

Decarlo, R.A., Saeks, R., and J. Murray, "A Nyquist-like Test for the Stability of Two-Dimensional Digital Filters", IEEE Proc., Vol. 65, pp. 978-979, (1977).

Asher, R., "Adaptive Estimation of Aberration Coefieients in Adaptive Optics", Information Sciences, Vol. 12, pp. 245-261, (1977).

Asher, R., "Perfect Decoupling of Linear Systems with Discrete Parameter Uncertainties", IEEE Trans. on Automatic Cont., Vol. AC-22, pp. 498-500, (1977).

Marks, R.J., Walkup, J., Hagler, M.O., and T.F. Krile, "Space Variant Processing of 1-D Signals", Appl. Optics, Vol. 16, pp. 739-745, (1977).

Hagler, M.O., Krile, T.F., Marks, R., and J. Walkup, "Holographic Representation of Space-Variant Systems Using Phase Coded Reference Beams", Appl. Optics, Vol. 16, (1977).

Portnoy, W.M., Emergency Medical Care, Lexington, Mass., Lexington Books, 1977.

Marks, R., Walkup, J., and T.F. Krile, "Ambiguity Function Display: An Improved Coherent Processor", Appl. Optics, Vol. 16, pp. 746-750, (1977).

Systems - Conference Papers and Reports

Saeks, R., and K.S. Chao, Proc. of the 20th Midwest Symposium on Circuits and Systems, North Hollywood, Western Periodicals Inc, 1977.

Saeks, R., and N. Levan, Proc. of the 2nd Int. Symp. on the Operator Theory of Networks and Systems, North Hollywood, Western Periodicals, Inc., 1977.

DeCarlo, R.A., Murray, J., and R. Saeks, "Three Graphical Tests for the Stability of Multidimensional Digital Filters", Proc. of the 1977 IEEE Int. Symp. on Circuits and Systems, Pheonix, 1977.

Leake, R.J., and R. Saeks, "On the Computability of Lacunary Sets", Proc. of the 1977 IEEE Int. Symp. on Circuits and Systems, Pheonix, 1977.

Sen, N., and R. Saeks, "A Measure of Testability and its Application to Test Point Selection - Computation", Proc. of the AUTOTESTCON '77, Hyannis, Mass., 1977.

Tavenner, D., Saeks R., and D. Gustafson, "Microprocessor Implementation of a Fault Prediction Algorithm", Proc. of the 20th Midwest Symp. on Circuits and Systems, Lubbock, Tx., 1977.

Sen, N., and R. Saeks, "A Measure of Testability and its Application to Test Point Selection - Theory", Proc. of the 20th Midwest Symp. on Circuits and Systems, Lubbock, TX,, 1977.

Tung, L., and R. Saeks, "An Experiment in Fault Prediction", Proc. of the 4th Symp. on Reliability in Electronics, Budapest, 1977.

Tung, L. and R. Saeks, "Wiener-Hopf Techniques in Resolution Space", Proc. of the 2nd Int. Symp. on the Operator Theory of Networks and Systems, Lubbock, Tx., 1977.

Asher, R., "Adaptive Estimation of Aberation Coeficients in Adaptive Optics", Proc. of the 1977 Joint Automatics Control Conf., San Francisco, 1977.

Asher, R., "Optimal and Sub-optimal Results in Full and Reduced Order Linear Filtering", Proc. of the 1977 Joint Automatic Control Conf., San Francisco, 1977.

Asher, R.B., "Adaptive Estimation of Samll Angle Measurements", Proc. of the AIAA Guidance and Cont. Conf., San Diego, 1977.

Asher, R., "Mixed Observable Estimation of Random Solar Electric Propulsion Spacecraft Thrusts", Proc. of the AIAA Guidance and Cont. Conf., San Diego, 1977.

Asher, R., "Fixed Configuration Filters for Image Processing", Proc. of the 1977 Int. Conf. on Cybernetics and Soc., Anahein, Ca., 1977.

Asher, R., "Optical Estimation in Signal Dependent Noise", Proc. of the Optical Soc. of Amer. Annual Meeting, San Diego, 1977.

Asher, R., "Adaptive Estimation of Phase Distortions in Adaptive Optics", Proc. of the Optical Soc. of Amer. Annual Meeting, San Diego, 1977.

Asher, R., "High Angle Attack Flight Cont. Using Stochastic Reference Model Adaptive Control", IEEE Decision and Cont. Conf., Orlando, Fla., 1977.

Asher, R., "Optimal and Sub-optimal Estimation of Mixed Rotational Observables", IEEE Decision and Cont. Conf., Orlando, Fla., 1977.

Chao, K.S., and D.K. Liu, "Transfer Characteristic Plots and Large-Cange Sensitivity of Nonlinear Resistive Networks", Proc. of the 1977 IEEE Int. Symp. on Circuits and Systems, Pheonix, 1977.

Pan, C.T. and K.S. Chao, "Difference Equation Approach for Solving Nonlinear Equations", Proc. of the 20th Midwest Symp. on Circuits and Systems, Lubbock, Tx, 1977.

Gustafson, D.L., "Review of Hughs' Mathematical Description of Linear Systems", Circuits and Systems, Vol. 10, p. 5, 1977.

Gustafson, D., "A Series of Microprocessor Courses", Digest of the Conf. of the ASEE Gulf-Southwest Section, Oklahoma City, 1977.

Prabhakar, J.C., "Synchronization Recovery", Proc. of the NATO Advanced Studies Workshop, London, 1977.

Prabhakar, J.C., "Walsh Transforms - Another Look", Proc. of the 20th Midwest Symp. on Circuits and Systems, Lubbock, TX., 1977.

Marks, R., Walkup, J., and M. Hagler, "Sampling Theorems for Linear Space-Variant Systems", Proc. of the 20th Midwest Symp. on Circuits and Systems, Lubbock, Tx., 1977.

Krile, T.F., Hagler, M., Marks, R., and J. Walkup, "Space Variant Holographic Optics Using Phase Coded Reference Beams", Proc. of the 1977 Int. Optical Comp. Conf., San Diego, 1977.

Bell, S.J., and M. Hagler, "Nyquist Stability Conditions for Laser Operational Amplifiers with Gaussian and Lorentizian Gain Properties", Proc. of the 20th Midwest Symp. on Circuits and Systems, Lubbock, Tx., 1977.

Hagler, M., Krile, T.F., Redis, W.D., and M.I. Jones, "Families of 2-Dimensional Phase-Coded Diffusers for Multiplex Holography", Proc. of the 1977 Meeting of the Optical Soc. of Amer., Toronto, 1977.

Froehlich, G., and J. Walkup, "Some Results on Detection and Estimation in Signal Dependent Noise", Proc. of the 20th Midwest Symp. on Circuits and Systems, Lubbock, Tx., 1977.

Marks, R., and J. Walkup, "Coherent Optical Processor for Ambiguity Function Display and One Dimensional Corrolation/Convolution Operations", SPIE Cong. on Optics in Radar, 1977.

Walkup, J., Froehlich, G., and R. Asher, "Optimal Estimation in Signal Dependent Noise", Proc. of the 1977 Meeting of the Optical Soc. of Amer., Toronto, 1977.

Walkup, J., Marks, R., and M. Irby, "Techniques in One Dimensional Space Variant Processing", Proc. of the 1977 Meeting of the Optical Soc. of Amer., 1977.

Systems - in Press

Gustafson, D., "Review of Peatman's Microcomputer Based Design", IEEE Spectrum, (to appear).

Hagler, M., Froehlich, G., and J. Walkup, "A Set of Optical Information Processing Experiments", IEEE Trans. on Education (to appear).

Marks, R., Walkup, J., and M. Hagler, "Sampling Theorems for Linear Time-Variant Systems", IEEE Trans. on Circuits and Systems, (to appear).

Physical Electronics - Refereed Publications

Williams, P.F., "Experimental Measurement of Dissociative Molecular Potential Functions from Continuous Resonance Raman Spectra", Chem. Physics Lett., Vol. 47, pp. 150-151, (1977).

Williams, P.F., "Evidence for Excited State Interference in Resonance Raman Scattering", Chem. Physics Lett., Vol. 50, pp. 57-59, (1977).

Bushnell, A.H., Gundersen, M., and T.R. Burkes, "Effect of a Small Capacitor in Parallel with a Pulsed CO_2 TEA Laser", IEEE Jour. on Quantum Elec., Vol. QE-12, pp. 447-449, (1977)

Gundersen, M., "Spectroscopy with Visable and Far-Infrared Lasers", Proc. of SPIE, Vol. 82, pp. 7-12, (1976).

Reichert, J.D., "Refraction", in Encyclopedia of Physics, Stroudsburg, Pa., Dowden, Hutchinson, and Ross, Inc., 1977.

Physical Electronics - Conference Papers and Reports

Gundersen, M., "Line Narrowing of CO_2 Lasers Required for Optical Pumping", Report AP-2-77-221, Los Alomos Scientific Laboratory, 1977.

Jones, C.R., Bushnell, A.H. and M. Gundersen, "An Optically Pumped $^{15}NH_3$ Laser", Proc. of the Conf. on Chemical and Molecular Lasers", St. Louis, 1977.

Gundersen, M., "Laser Research", Report of the Conf. on Lasers for Isotope Separation", Albuquerque, 1977.

Gundersen, M., "Spectroscopy with Visible and Far-Infrared Lasers", Proc. of the SPIE Conf., 1976.

Physical Electronics - in Press

Williams, P.F., and D. Russou, "The Resonance Raman Effect", in Topics in Appl. Physics, (to appear).

Jones, C.R., Buchwald, M.I., Gundersen, M., and A.H. Bushnell, "Ammonia Laser Optically Pumped with an HF Laser", Optics Communications, (to appear).

Electromagnetics - Refereed Publications

Hagler, M., and M. Kristiansen, "A Numerical Analysis of High Power Laser Propagation in Magnetized Plasma", in Recent Advances in Plasma Physics, Madras, Indian Acad. of Sci., 1977.

Kunhardt, E., "Propagation of Nonlinear Waves Along Magneto Plasma Columns", Phys. of Fluids, Vol. 20, pp. 1499-1508, (1977).

Kunhardt, E., "Rotating Ball Generator", Rev. Sci. Inst., Dec. 1977.

Electromagnetics - Conference Papers and Reports

Hagler, M., and M. Kristiansen, "Ablation Rates of Sphereical Pellets in a Theta Pinch", IEEE Int. Conf. on Plasma Science, Troy, N.Y., 1977.

Kristiansen, M., and M. Hagler, "An Experimental Arrangement for Laser Beat Heating of Plasma", Proc. of the Engrg. Problems of Fusion Research Conf., Knoxville, 1977.

Hagler, M., and M. Kristiansen, "Description of the Texas Tech Tokomak", APS Plasma Physics Div. Meeting, Atlanta, 1977.

Trost, T., "Radio and Electric Fields Change Measurements on Severe Thunderstorms", Digest of the Amer. Geophysical Union Meeting, 1977.

Electromagnetics - in Press

Kristiansen, M., and M. Hagler, "Ablation Rates for Polystyrene Microspheres in a Theta Pinch Plasma", Jour. of Appl. Phys, (to appear).

Power - Refereed Publications

Kristiansen, M., and M. Hagler, An Introduction to Controlled Thermonuclear Fusion, Lexington Press, Lexington, Mass., 1977.

Power - Conference Papers and Reports

Burkes, T.R. and J.P. Craig, "Power System Analysis", AFWL Tech. Report, Kirtland AFB, N.M., 1977.

Craig, J.P., and S. Tsai, "A Variable Speed Wind Generator", Texas Power and Light Co., Tech Report, Dallas, Tx., 1977.

Craig, J.P., "Evaluation of Projected Effectiveness of the National Energy Plan", Texas Tech Univ. Center for Energy Research Tech. Report, Lubbock, Tx., 1977.

Trost, T., "Pulse Power System Employing Magnetic Energy Storage", Naval Surface Weapons Center Tech. Report, Dahlgren, Va., 1977.

Reichert, J.D., "A Strategy of Calculation of Optical Concentration Distributions for Fixed Mirror Systems", Proc. of the ERDA Solar Energy Workshop on Methods for Optical Mirror Systems, Livermore, Calif., 1977.

Reichert, J.D., "The Crosbyton Solor Power Project: Fixed Sphereical Mirror/Tracking Receiver", Proc. of the ERDA Concentrating Collector Conf., Washington, D.C., 1977,